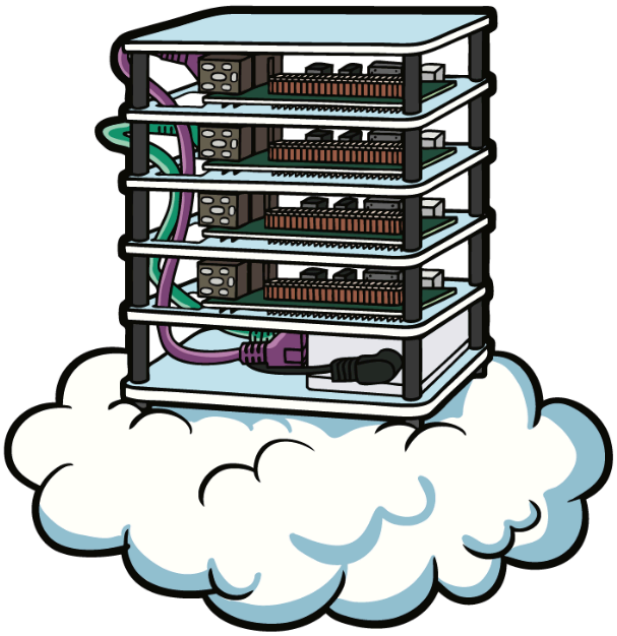
Anàlisi de requisits i especificació d'una BD per a l'estudi d'impactes publicitaris

Treball de Final de Grau (2 de juliol de 2018)
Enginyeria del Software



Autor: Jordi Estapé Canal

Director: Sandra Rodríguez Borau

Ponent: Joan Antoni Pastor Collado (ESSI)

Empresa: IKI Media Communications S.L.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

FIB

Abstract

Català

En aquest treball s'ha buscat trobar una solució software basada en una base de dades relacional per a l'emmagatzemament i l'encreuament de les diferents fonts de dades usades en l'estudi analític de l'impacte publicitàri.

Els principals objectius del projecte han estat la realització d'un estudi de context (per a entendre el que s'ha de solucionar i el context en el qual ha d'encaixar la solució), l'especificació de requisits (per a determinar les necessitats del sistema a dissenyar), el disseny d'una solució que asseguri que tant els requisits es compleixen com el seu encaix dins del context i la demostració de viabilitat de la solució.

Per a l'estudi de context, s'ha portat a terme una anàlisi de les fonts de dades existents, l'estudi dels softwares i els hardwares existents, una anàlisi dels stakeholders i un estudi del flux d'actuació actual.

Pel que fa a l'especificació de requisits s'ha realitzat una primera selecció d'aquests a partir de l'estudi de context i posteriorment la seva traducció a casos d'ús.

Seguidament i tenint en compte tant els requisits com el context estudiat s'ha dissenyat una solució mitjançant el disseny de la base de dades, el disseny del flux d'execució del domini i el disseny la navegació in-app i del front-end.

Finalment s'ha proposat un flux de procés que demostra el funcionament i integració de la proposta en el cas d'estudi específic i s'ha implementat un prototip amb l'objectiu de mostrar la solució dels punts crítics del projecte.

S'ha conclos el projecte donant una solució vàlida i assolint tots els objectius plantejats. A més a més, s'ha ofert un prototip de la solució proposada (fet que no havia estat contemplat en els objectius inicials).

Castellano

En este trabajo se ha buscado encontrar una solución software basada en una base de datos relacional para el almacenamiento y el cruce de las diferentes fuentes de datos usadas en el estudio analítico del impacto publicitario.

Los principales objetivos del proyecto han sido la realización de un estudio de contexto (para entender lo que se debe solucionar y el contexto en el que tiene que encajar la solución), la especificación de requisitos (para determinar las necesidades del sistema a diseñar), el diseño de una solución que asegure que tanto los requisitos se cumplen como su encaje dentro del contexto y la demostración de viabilidad de la solución.

Para el estudio de contexto, se ha llevado a cabo un análisis de las fuentes de datos existentes, un estudio

de los softwares y los hardwares existentes, un análisis de los stakeholders y un estudio del flujo de actuación actual.

En cuanto a la especificación de requisitos se ha realizado una primera selección de éstos a partir del estudio de contexto y posteriormente se han traducido a casos de uso.

Seguidamente y teniendo en cuenta tanto los requisitos como el contexto estudiado se ha diseñado una solución mediante el diseño de la base de datos, el diseño del flujo de ejecución del dominio y el diseño de la navegación in-app y del front-end.

Finalmente se ha propuesto un flujo de proceso que demuestra el funcionamiento e integración de la propuesta en el caso de estudio específico y se ha implementado un prototipo con el objetivo de mostrar la solución de los puntos críticos del proyecto.

Se ha concluido el proyecto dando una solución válida y logrando todos los objetivos planteados. Además, se ha ofrecido un prototipo de la solución propuesta (cosa que no había sido contemplado en los objetivos iniciales).

English

In this work we have sought to find a software solution based on a relational database for the storage and crossing of the different data sources used in the analytical study of advertising impact.

The main objectives of the project have been to carry out a context study (to understand what needs to be resolved and the context in which the solution should fit), the specification of requirements (to determine the needs of the system to be designed), the design of a solution that ensures that the requirements are met and that its integration within the context is valid. The last objective is to proof the solution viability by demonstrating how its usage may improve actual results.

For the context study, an analysis of the existing data sources, the study of existing software and hardware, an analysis of the stakeholders and a study of the current flow of process has been carried out.

Regarding the specification of requirements, a first selection based on the context study has been carried out and after that, it translated to use cases diagram.

Then, taking into account both the requirements and the context studied, a solution has been designed through the design of the database, the design of the flow of execution for the domain and the design of the front-end navigation.

Finally, a process flow that demonstrates the operation and integration of the proposal in the case of this specific study has been proposed and a prototype has been implemented in order to show the solution to the critical points of the project.

The project has been concluded giving a valid solution and achieving all the objectives set. In addition, a prototype of the proposed solution has been offered (fact that had not been contemplated in the initial objectives).

Agraïments

A la Sandra Rodriguez Borau, per a la seva brillant direcció d'aquest projecte. Per la seva disposició a indicar-me i a facilitar que la realització no es realentitzes. Gràcies per treure temps d'on no n'hi havia i per la teva incansable cooperació activa en la realització del projecte.

A en Joan Antoni Pastor, per la seva excel·lent gestió i ajuda en aquest projecte des de la figura de Ponent. Per haver estat un guia en aquest camí ple d'entrebancs i per aconsellar-me en tots aquells detalls que milloraven el resultat final del projecte. Moltes gràcies per la teva disposició, gairebé a un dia vista, i per la teva relació propera amb el projecte.

A l'equip d'IKI Media Communications S.L. per haver-me tractat com un més de l'empresa i per l'esforç realitzat per totes les parts per trobar espais en els horaris per a totes les reunions que se'ls ha demanat.

A la meva parella, Ariadna Boada, per la paciència que ha demostrat envers els meus nervis davant la càrrega de feina ha realitzat i per la seva col·laboració en la revisió final del treball.

I finalment però no menys important, a tota la meva família pel suport moral que m'han ofert i per acompanyar-me durant la realització d'aquest projecte.

Índex

Índex de figures

Índex de taules

1

Introducció

El valor afegit que dóna l'estudi analític de les dades en el món publicitari està revolucionant el sector: el fet de poder realitzar prediccions de l'efecte que les campanyes produiran, passarà, en un futur no gaire llunyà, de ser un valor afegit a quelcom necessari per a la justificació de les decisions de planning de les campanyes i la millora de resultats d'aquestes.

A causa de la instauració profunda de la data science en el món publicitari, ens trobem en una situació en la qual empreses importants del sector dediquen gran part de les seves inversions a créixer en l'àmbit del tractament i l'estudi de les dades per encabir-se en aquest mercat emergent. Un molt bon exemple d'aquest fet és WWP (considerada l'agencia més gran del món segons 'Advertising Age'), que inicialment era un conglomerat d'empreses purament publicitaries, en aquest moment està obrint noves vies envers l'estudi analític i la consultoria fent que actualment no es consideri un grup principalment dedicat a la publicitat. Veiem també com Accenture, IBM i Deloitte (totes dedicades a consultoria IT) apareixen també en bones posicions a la llista, quan fa 5 anys ni tan sols hi apareixen. [3]

Queda clar doncs el canvi de la professió en els últims anys a causa de l'aparició del Big Data i el pes que ha resultat agafar l'estudi analític de les dades en el sector. Aquest fet ha portat l'empresa IKI Media Communications S.L. a fer un gir cap aquest àmbit de treball per a poder ajustar-se a les noves necessitats establertes pel client.

Però, al ser una mitjana empresa amb tan sols 3 anys de vida, les dificultats per a mantenir i dur a terme aquest servei han estat immenses, ja que, a diferència d'altres empreses, no es comptava amb una inversió inicial gran per al desenvolupament d'eines o sistemes que facilitessin i agilitzessin el procés d'estudi de les dades. A més el temps per a dedicar al desenvolupament d'eines o la millora de procés del departament de Data Analytics tampoc va resultar suficient a causa del volum de feina que és va haver d'afrontar. Per tant, ens trobàvem davant un departament que estava realitzant tasques diàries d'anàlisi de dades de manera manual, amb les eines mínimes i sense definir clarament un procés (per culpa de la falta de temps i material). Podem afirmar però, que actualment gràcies a l'esforç realitzat pels treballadors, el departament ha tirat endavant i actualment es troba en un moment més estable i preparat per a créixer tecnològicament per a poder assumir càrregues de feina superiors i reduir tant els riscos derivats del factor humà com el temps de realització de cada estudi. Després d'una anàlisi inicial de la situació del departament s'han detectat problemes amb l'extracció i l'encreuament de les dades. A més a més s'ha detectat que el flux d'actuació sobre els datasets no està definit.

Partint de la situació inicial plantejada, en aquest projecte de final de grau es tractarà de trobar una solució software que permeti resoldre els problemes d'extracció i encreuament de les dades i de preprocessing d'aquestes. És a dir, es pretén realitzar el disseny d'un sistema de base de dades que permeti unificar tota la informació rebuda de manera automàtica i el plantejament d'un mètode de lectura i preprocessing

de les dades que permeti mantenir aquest sistema en funcionament i actualitzat. Idíl·licament aquest sistema de preprocessament de les dades hauria d'automatitzar-se gairebé completament i hauria de ser prou canviable per a ajustar-se a les modificacions que cada client demani.

2

Estudi Inicial de Context

2.1 Context específic

Comencem ara amb l'estudi del context específic del cas d'estudi. En aquest punt, un cop explicada la necessitat real de l'anàlisi de dades en el món de la publicitat (1 - Introducció), es pretén plantejar la situació actual de l'empresa en major profunditat en aquest àmbit per tal de situar el lector en el punt d'inici des del qual es parteix amb la intenció d'identificar els principals problemes que s'hauran d'afrontar i definir-los acuradament. Per a situar el cas d'estudi presentem la definició que IKI Media Communications S.L. presenta sobre si mateixa:

> **Citació 1** *'IKI es una nueva agencia de medios y comunicación en la que trabajamos en colaboración con nuestros clientes, para alcanzar soluciones de comunicación claras e inteligentes y que se adapten a los objetivos de negocio de los anunciantes.' [4]*

Queda clar doncs que l'empresa IKI Media Communications S.L. fa d'intermediària en el procés de compra d'espais publicitaris i que a part de realitzar aquesta funció, ofereix adaptació als interessos del client i solucions específiques per cada un d'ells. Aquesta última oferta és la que fa que l'empresa tingui un diferencial respecte la competència, ja que, la realització d'aquestes estratègies no és tan sols basada en l'experiència de l'equip sinó que s'ha creat un departament de Data Science amb l'objectiu d'oferir una justificació de la planificació de cada campanya basada en estudis reals dels diferents escenaris possibles.

Recordem també que IKI Media Communications S.L. tan sols té 3 anys de recorregut tot i que compte amb personal qualificat i amb experiència en la gestió d'aquest tipus de negoci en el sector. Cal remarcar però que l'experiència prèvia quant a l'estudi analític de modelatge era gairebé nul·la. Ja s'ha dit al punt d'introducció (1 - Introducció) que la entrada del Data Science dins del sector ha succeït durant els últims 5 anys i per tant la situació pel que fa a experiència és totalment justificable i encara més quan ens trobem davant una empresa mitjana en plena fase de creixement. Trobem de vital importància tenir en compte aquest fet quan s'estudia el cas, ja que, condiciona plenament tant l'estat actual com les condicions del sistema a desenvolupar.

Un cop contextualitzada l'empresa i l'existència del departament passem a basar-nos únicament en el que afectarà directament al cas d'estudi, és a dir, al grup de Data Science. Aquest equip està format per 4 persones (1 analista programador, 2 estadístics i 1 enginyer del software) que treballen conjuntament per a la realització de les anàlisis. La funció de cada un d'ells es troba teòricament definida com s'explica a continuació:

- **Analista programador:** dedicat a l'extracció i tractament de les dades per a fer-les llegibles.
- **Enginyer del software:** dedicat a l'encreuament i la preparació dels conjunts de dades per les anàlisis.
- **Estadístics:** dedicats a la realització d'anàlisis i la selecció de dades i factors d'estudi. També realitzen les presentacions de resultats als clients.

Tot i trobar-nos amb una divisió de tasques teòrica, s'ha de tenir en compte que en casos normals, aquesta divisió de tasques resulta inviable doncs el procés d'extracció, tractament i encreuament de les dades resulta ser molt lent a causa dels constants canvis de formats i a l'ús de mètodes rudimentaris en comptes de softwares o sistemes especialitats. Per tant, aquí trobem un dels principals problemes, que és que les dades que s'extreuen i es llegeixen no són tractades de manera eficient i això genera un coll d'ampolla de cara al procés d'anàlisi. A més com s'ha dit, el mètode de tractament és rudimentari, fet que, incrementa el risc d'error humà. A tot això també cal sumar-li què, per culpa del coll d'ampolla i com ha estat explicat anteriorment, les tasques passen a ser realitzada per persones a les quals no haurien d'estar assignades, cosa que, fa que l'estandardització d'un procés d'actuació sobre les dades sigui difícil de determinar en el moment actual pel fet que, en molts casos aquest procés depèn totalment de l'estat de càrrega de feina del grup fent impossible l'establiment d'un ordre en les tasques i sovint trencant vincles de dependència entre tasques.

Un altre punt a tenir en compte és que actualment gràcies a l'experiència que s'ha guanyat durant aquests anys, s'ha establert una metodologia de treball que, tot i no aplicar-se en tots els casos ens pot donar una idea de com s'hauria de realitzar el procés i pot facilitar la concepció d'un sistema que permeti actuar amb un flux de treball similar al que actualment es troba funcionant per a facilitar l'adaptació de l'equip al nou sistema i disminuir la resistència al canvi que totes les parts apliquen inconscientment o conscientment.

Finalment, una vegada explicats els principals problemes als quals s'enfronta actualment el departament, plantegem quines són les intencions de l'equip per a millorar aquesta situació. Primerament la seva intenció és reduir el temps de lectura, tractament i encreuament de dades mitjançant un sistema que permeti agilitzar-ho. Haurà de ser condició d'aquest sistema que encaixi amb tots els softwares en ús i què sigui molt canviable. Es pretén també trobar un sistema que permeti mantenir un històric de les dades recollides per a tenir una base més gran sobre la qual treballar les anàlisis i també per poder començar a usar softwares com Tableau per donar als clients un seguiment més freqüent de les dades què es tracten i de l'impacte publicitari que s'està tenint. També tenen com a objectiu establir un funcionament que permeti millorar l'eficiència de treball actual i que els permeti realitzar anàlisis més freqüents i més complerts.

Concloem doncs que l'empresa IKI Media Communications i en especial el departament de Data Science de la mateixa té intenció de créixer i millorar el seu servei mitjançant la implantació d'un sistema informàtic que permeti tractar i gestionar les dades permeten extreure conjunts de dades preparats per a treballar de manera més ràpida. També val la pena afegir que a part d'aquest objectiu en tenen d'altres que podríem considerar secundaris com serien l'emmagatzemament de les dades en un sistema que permeti l'accés d'altres softwares i la millora de procés per a incrementar l'eficiència i reduir el risc.

2.2 Stakeholders

La realització d'aquest treball per tant tindrà un efecte directe sobre un conjunt de persones. Per tal de trobar el màxim nombre d'entitats afectades s'ha realitzat un estudi dels diferents stakeholders inicial que posteriorment serà estès per a una major comprensió dels stakeholders realment rellevants. S'han classificat els stakeholders detectats segons el seu posicionament (intern o extern dins l'empresa) i queden representats en la següent gràfica (Figura 2.1):

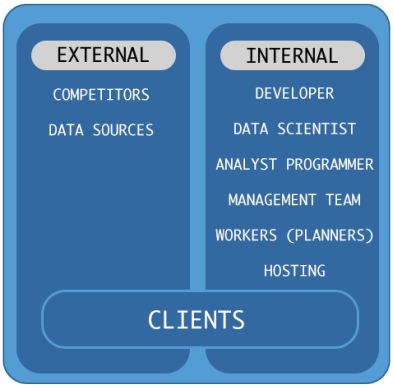


Figura 2.1: Identificació dels Stakeholders

Ara, una vegada enumerats els principals stakeholders del projecte, passem a especificar la descripció individual reduïda dels que resulten suficientment rellevants per a l'estudi. Per la realització d'aquesta anàlisi, s'ha decidit generar una estructura bàsica a manera de plantilla que permeti omplir les dades de cada un dels stakeholders obtenint així una descripció completa de cada un d'ells [5].

2.2.1 Competidors

- **Descripció** Són el conjunt d'empreses mitjanes del sector publicitari que fan una funció d'intermediari entre el client i el servei on apareixerà l'anunci, campanya o promoció.
- **Benefici o Pèrdua** Si el sistema funciona, hauran de millorar el seu sistema per tal d'assumir el mateix ritme de treball que l'empresa IKI Media Communications S.L. i també per tal d'oferir un servei similar. (Pèrdua)
- **Motivació** Inexistent, de fet no els afavoreix.
- **Que vol** Mantenir el mateix servei o ser ells els que innoven i treuen serveis diferencials respecte d'altres empreses del sector.
- **Com comunicar-s'hi** En principi no cal mantenir comunicació amb la competència.
- **Estat actual** Moltes de les empreses equiparables a IKI Media Communications S.L. no tenen servei de 'Data Science', per tant, es troben un pas per darrere i es troben en un entorn còmode amb competitivitat basada en marges de benefici i pricing.
- **Com superar estat actual?** Si es pot portar a terme el projecte i es normalitza el servei, passarà de ser un diferencial a ser quelcom necessari i es revaloraran els estudis analítics en la publicitat.
- **Influenciat per** El mercat i els clients.

2.2.2 Fonts de dades

- **Descripció** Conjunt d'empreses dedicades a la recollida i distribució de dades dels diferents mitjans en els quals es realitza publicitat (Kantar Media, Infoadex...)
- **Benefici o Pèrdua** Es facilitarà l'ús de les dades i probablement es realitzaran extensions de llicències per a obtenir més dades, ja que, en reduir el temps de treball en preprocessing es podrà analitzar un nombre superior de factors. (BENEFICI)
- **Motivació** Incrementar les llicencies tant en nombre com en nivell i donar valor al seu producte.
- **Que vol** Millorar l'experiència d'usuari i incrementar les seves vendes de llicències.
- **Com comunicar-s'hi** Mitjançant e-mail i tan sols per demanar els fitxers de metadades.
- **Estat actual** Empreses de gran reconeixement i fiables. Gran quantitat de dades coherents i útils. Els usuaris (D'IKI Media Communications S.L.) no n'usen tot el potencial.
- **Com superar estat actual?** Trobar un mètode el màxim d'automatitzat possible per a reduir el temps de preprocessing i encreuament donant l'oportunitat als treballadors de treballar amb més factors i realitzar anàlisis més complets.
- **Influenciat per** El mercat i les agències de mitjans.

2.2.3 Desenvolupador

- **Descripció** El desenvolupador del futur sistema, que en aquest cas seré jo.
- **Benefici o Pèrdua** Millora del coneixement i possible millora de la posició laboral en cas d'aconseguir un sistema vàlid i usable.
- **Motivació** Finalització del TFG i interès personal dins l'empresa.
- **Que vol** Realitzar un bon sistema software que permeti millorar el sistema actual i millorar posició laboral dins l'empresa.
- **Com comunicar-s'hi** -.
- **Estat actual** Intentant comprendre tots els factors, estudiant el context del problema plantejat i realitzant ajudes a l'equip en preprocessament i encreuament.
- **Com superar estat actual?** Mitjançant consultes amb el director del TFG (Sandra Rodríguez) i amb els companys de departament per acabar d'entendre el context i trobar un sistema software que hi encaixi.
- **Influenciat per** Director, ponent, companys de departament i equip directiu de l'empresa.

2.2.4 Estadístics

- **Descripció** Persona que realitza les tasques de modelització i estudi de les dades. També col·labora en la funció de preprocessament i encreuament d'aquestes.
- **Benefici o Pèrdua** Deixar de realitzar la funció de preprocessament i encreuament de dades, obtenint més temps per a la realització de les tasques de modelització i reduint també carrega de treball. (BENEFICI)
- **Motivació** Reduir la càrrega de treball i tenir més temps per a la realització d'estudis més complets.
- **Que vol** Tenir les dades preparades per a treballar-les i haver-les de tractar el mínim possible.
- **Com comunicar-s'hi** Mitjançant reunions en persona i e-mails.

- **Estat actual** Sobrecarrega de feina i amb poques esperances de millora mitjançant un sistema de bases de dades per males experiències prèvies en altres empreses.
- **Com superar estat actual?** Oferir un software que permeti reduir la seva càrrega de feina i anar donant resultats a poc a poc per tal de millorar la visió dels sistemes de bases de dades.
- **Influenciat per** Equip directiu de l'empresa i analista programador.

2.2.5 Analista programador

- **Descripció** Persona dedicada al preprocessament de dades i gestió de les lectures dels datasources. Especialista en Excel i desenvolupador de gairebé tot el software de l'empresa.
- **Benefici o Pèrdua** Reduir la seva càrrega de treball i fer-la més simple en quant a preprocessament i gestió de les dades. (BENEFICI)
- **Motivació** Reduir la seva càrrega de treball i tenir més temps per a la implementació de nous softwares per l'empresa.
- **Que vol** Tenir un sistema de lectura de dades i preprocessament automàtic que faciliti la seva tasca.
- **Com comunicar-s'hi** Mitjançant reunions en persona i e-mails.
- **Estat actual** Sobrecarrega de feina. Poc temps i pocs recursos per al desenvolupament de noves eines.
- **Com superar estat actual?** Oferir un software que permeti reduir la seva càrrega de feina i documentació per tal de donar l'opció d'usar el sistema dissenyat com a datasource del software de l'empresa.
- **Influenciat per** Equip directiu de l'empresa i analista de dades.

2.2.6 Equip directiu

- **Descripció** Conjunt de persones que prenen les decisions i gestionen la inversió de l'empresa. En determinen el futur.
- **Benefici o Pèrdua** Millorar el servei que la seva empresa ofereix. (BENEFICI)
- **Motivació** Incrementar el valor de l'empresa i millorar el servei que dóna la mateixa.
- **Que vol** Incrementar el ritme de producció d'anàlisis i la qualitat dels mateixos sense incrementar el nombre de treballadors.
- **Com comunicar-s'hi** Mitjançant reunions en persona i e-mails. Com que el departament de data science es troba molt ben considerat, es manté un contacte proper amb l'equip directiu.
- **Estat actual** Treballant per fer créixer l'empresa i augmentar la facturació i els marges.
- **Com superar estat actual?** Oferir un software que redueixi el temps de realització de les tasques d'anàlisi i que redueixi el risc d'error durant aquestes.
- **Influenciat per** Mercat i client.

2.2.7 Client

- **Descripció** És l'entitat que contracta els serveis d'IKI Media Communications S.L.
- **Benefici o Pèrdua** Podrà rebre estudis complets amb més freqüència, més extensos i amb menys errors.
- **Motivació** Rebre un millor servei.
- **Que vol** Rebre el millor servei i tenir unes anàlisis completes que els hi permetin prendre decisions sobre les inversions publicitaries de l'entitat.
- **Com comunicar-s'hi** Mitjançant e-mail o trucada.
- **Estat actual** Rebent anàlisi mensualment i amb reticència a enviar dades.
- **Com superar estat actual?** Un cop el sistema funcioni, donar resultats amb més freqüència i guanyar confiança gràcies a prediccions i modelatges encertats.
- **Influenciat per** Equip directiu.

2.2.8 Prioritització Poder vs Interes dels Stakeholders

Un cop realitzat aquest estudi individual dels diferents stakeholders tenim una visió general de les necessitats i condicions de cada un d'ells, fet que ens ajudarà en un futur a realitzar una millor presa de decisions. Un altre punt que realitzarem serà la classificació dels stakeholders (segons: 'Environmental Scanning - The Impact of the Stakeholder Concept' [6]) per tal de poder saber el tipus d'interacció i la prioritat d'aquests. La gràfica obtinguda és la següent (Figura 2.2):

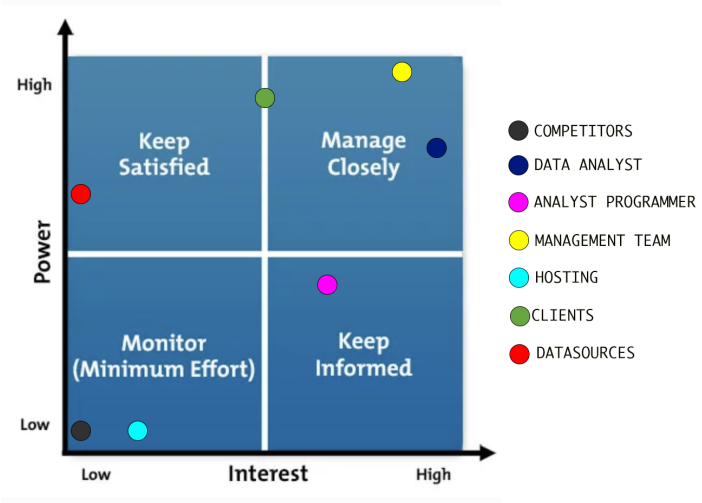


Figura 2.2: Reixeta per a la prioritització dels stakeholders: Poder vs. Interés

Veiem doncs amb aquesta gràfica com s'ha de tractar cada un dels stakeholders que s'han estudiat anteriorment i contemplem que ens hem de basar en tenir un bon tracte i mantenir informats als que considerem stakeholders interns (Figura 2.1) que són els que més impacte tenen sobre el projecte.

2.3 Estat de l'art

En l'actualitat l'obtenció i preparació sol realitzar-se mitjançant processos iteratius de tractament de les dades [7]. Identifiquem 3 àmbits a estudiar quan parlem d'estat de l'art el primer relacionat amb l'extracció de dades, posteriorment el tractament d'aquestes, altrament anomenat 'preprocessing' i finalment els softwares existents per a la realització tant de l'extracció com del tractament de dades. Tractarem doncs d'estudiar aquests tres àmbits per separat.

2.3.1 Extracció de dades

El procés de recollida de dades va plenament lligat a les fonts de dades de les quals s'extreu la informació a treballar. Aquestes fonts de dades hauran de ser seleccionades tenint en compte el problema plantejat, la solució o anàlisi que es vol donar (el resultat a obtenir) i les dades existents, valorant així quines fonts de dades aporten informació rellevant per al cas d'estudi proposat [1]. Així doncs i com podem observar a la Figura 2.3, serà la interacció dels tres factors prèviament seleccionats la que ens facilitarà la selecció de totes les dades a usar.

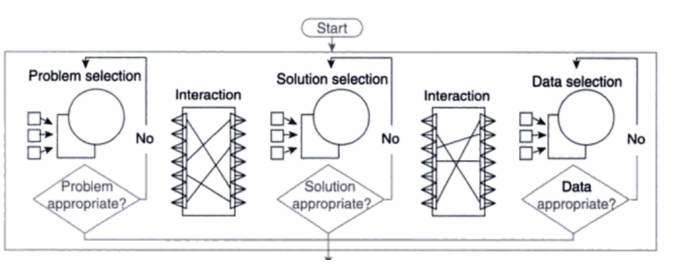


Figura 2.3: Factors rellevants de l'extracció de dades [1]

Aquesta selecció haurà de ser rigorosa, perquè qualsevol modelatge amb dades alterades pot arribar a desviar el resultat allunyant-lo de la realitat. Cal donar doncs especial rellevància a la verificació i auditació de la qualitat de les dades [8]. Aquesta qualitat sovint és mesurada amb índexs relacionats amb el cas d'estudi, aquests índexs són anomenats 'Case-based reasoning' (CDR) i són sistemes basats en l'estimació de validesa d'un cas base, fet que posteriorment permet fer un increment en el nombre de dades, doncs si l'estimació de les dades del cas base segueixen essent correcte, la qualitat general de les dades també es mantindrà. Alguns criteris a tenir en compte quan parlem d'aquest índex a crear són:

1. *Accessibilitat.* El grau en què es troben les dades disponibles, o fàcilment i ràpidament recuperables.
2. *Quantitat adequada de dades.* El volum de dades és adequat pel cas d'estudi.
3. *Credibilitat.* Mesura de fins a quin punt les dades es consideren veritables i creïbles.
4. *Completesa.* Grau en què les dades es troben completes (no N/A) i són suficients per a l'estudi.
5. *Representació concreta.* En quina mesura les dades estan representades de forma compacta.
6. *Representació coherent.* Fins a quin punt les dades es presenten en el mateix format.
7. *Facilitat de manipulació.* El grau en què les dades són fàcils de manipular i aplicar a diferents tasques.

8. *Llibertat d'errors.* Mesura la correctesa de les dades i la seva fiabilitat.
9. *Interpretació.* El grau d'adequació de les dades a idiomes, símbols i unitats. Les definicions de metadades són clares.
10. *Objectivitat.* Fins a quin punt les dades són imparcials, sense prejudicis.
11. *Rellevància.* En quina mesura les dades són útils i rellevants per al cas d'estudi. Aporten valor?
12. *Reputació.* Fins a quin punt les dades són valorades pel que fa a la seva font o contingut.
13. *Seguretat.* Mesura que el grau d'accés a les dades sigui restringit adequadament per mantenir la seva seguretat.
14. *Puntualitat.* Fins a quin punt les dades estan prou actualitzades per a la tasca que es presenta.
15. *Comprensibilitat.* Nivell comprensió de les dades.
16. *Valor afegit.* Grau de valor que afegeixen les dades seleccionades.

A partir de la selecció dels índexs aplicables al cas d'estudi es realitza un subindex (Idx) per a cada cas i se li assigna un pes (W). Finalment es realitza el càlcul de l'índex general mitjançant la fórmula (Figura 2.4):

$$\frac{\hat{I}dx_I \cdot \text{Weight}_I + Idx_{II} \cdot \text{Weight}_{II} + Idx_{III} \cdot \text{Weight}_{III}}{\sum_{i=1}^{i=3} \text{Weight}_i}$$

Figura 2.4: Formula Índex Qualitat Dades (cas de 3 subindexos) [2]

Obtenim per tant d'aquest mètode un valor que ens permetrà estudiar la qualitat de les dades i establir un llindar a superar per assegurar la qualitat d'aquestes. [2].

A més a més d'aquests mètodes d'avaluació de la qualitat resulta important donar informació sobre la principal tendència actual quant a tractament i anàlisi de dades anomenada Knowledge Discovery in Databases (KDD) que es refereix a ün procés no trivial d'identificar patrons vàlids, nous, potencialment útils i comprensibles en les dades"[9]. L'extracció de dades es troba en el primer dels passos i acaba determinant la informació a tractar. Es realitzarà una anàlisi més exhaustiu de KDD i d'altres Data Mining Processes (DPM) en el següent punt ja aporten més informació sobre preprocessing que sobre l'obtenció de les dades en si.

2.3.2 Preprocessing

El món del Data Mining està avançant molt en les últimes dècades a causa de l'increment d'interès en el Machine Learning, les IA, els estudis predictius i de modelatge de marketing, etc. Estan apareixent artefactes anomenats 'Data Mining Process' (DMP) que volen modelitzar principalment el tractament de les dades a més a més del procés d'anàlisi dels mateixos que queda fora del nostre àmbit de treball [10]. Cal remarcar que aquests DMP pretenen estandarditzar els processos assegurant la qualitat i rigorositat dels resultats però que en cap moment automatitzen els diferents processos a realitzar, perquè això és quelcom que cada entitat que usi un DMP haurà d'ajustar i implementar en dependència amb les seves necessitats. Alguns dels més interessants pel nostre cas d'estudi són el KKD Process [11] i el Cross Industry Standard Process for Data Mining (CRISP-DM) [12] que s'explicaran a continuació.

Comencem doncs per l'explicació del KDD. I presentant el flux de processos que proposa, observables en la Figura 2.5.

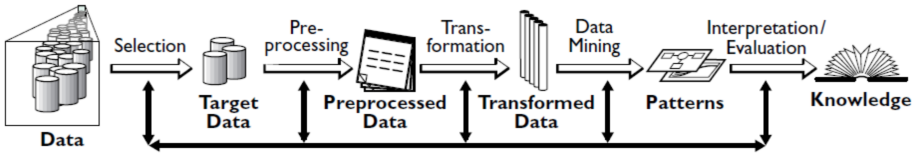


Figura 2.5: Diagrama del KDD Process

Veiem doncs que KDD ens proposa un sistema de flux basat en 5 processos i 6 elements. Identifiquem també que, pel que fa al nostre projecte, es limita a tractar els dos primers processos i per tant tan sols els tres primers elements, tot i que, de cara a millorar-ne la usabilitat també podríem incloure la tercera fase afegint doncs també un quart element. Realitzem ara un estudi dels diferents processos rellevants pel nostre scope [13]:

1. *Creació d'un conjunt de dades.* Inclou seleccionar un conjunt de dades o centrar-se en un subconjunt de variables i analitzar-ne la validesa, formant doncs un conjunt de dades de qualitat per a l'anàlisi.

2. *Neteja de les dades i preprocessament.* Inclou operacions bàsiques, com ara l'eliminació de sorolls o 'outliers', encreuament de les dades, preparació de la informació necessària per modelitzar, solucions a la 'missing data', decisió de tipus de camps, etc. Bàsicament es tracta d'un pas en el qual es tracten les dades per, a partir d'un conjunt de dades brutes obtingudes en la primera fase, generar un conjunt nét de dades, sobre el que es pugui realitzar l'anàlisi.

3. *Reducció o projecció de les dades.* Es refereix a un procés purament estadístic en el que s'intenta reduir la dimensionalitat de les dades o trobar representacions alternatives invariants d'aquestes.

Una vegada estudiats els processos de KDD ens adonem que, tot i que és ben cert que ens aporten una base cap a la realització del flux de processos de treball, no tindran un impacte directe de cara al disseny de la base de dades i la solució a proposar pel que fa a software i hardware. Com prèviament s'ha explicitat, pretenen tan sols donar un model de treball per a l'estandarditzadió dels processos i l'assegurança d'uns resultats correctes, i per tant, no n'asseguren l'automatització o fusió amb la tecnologia.

Passem doncs ara a l'estudi del mètode CRISP-DM per veure fins a quin punt ens pot aportar una milloria respecte el KDD. Com anteriorment en el cas de KDD comencem presentat el diagrama de flux que proposa CRISP-DM.

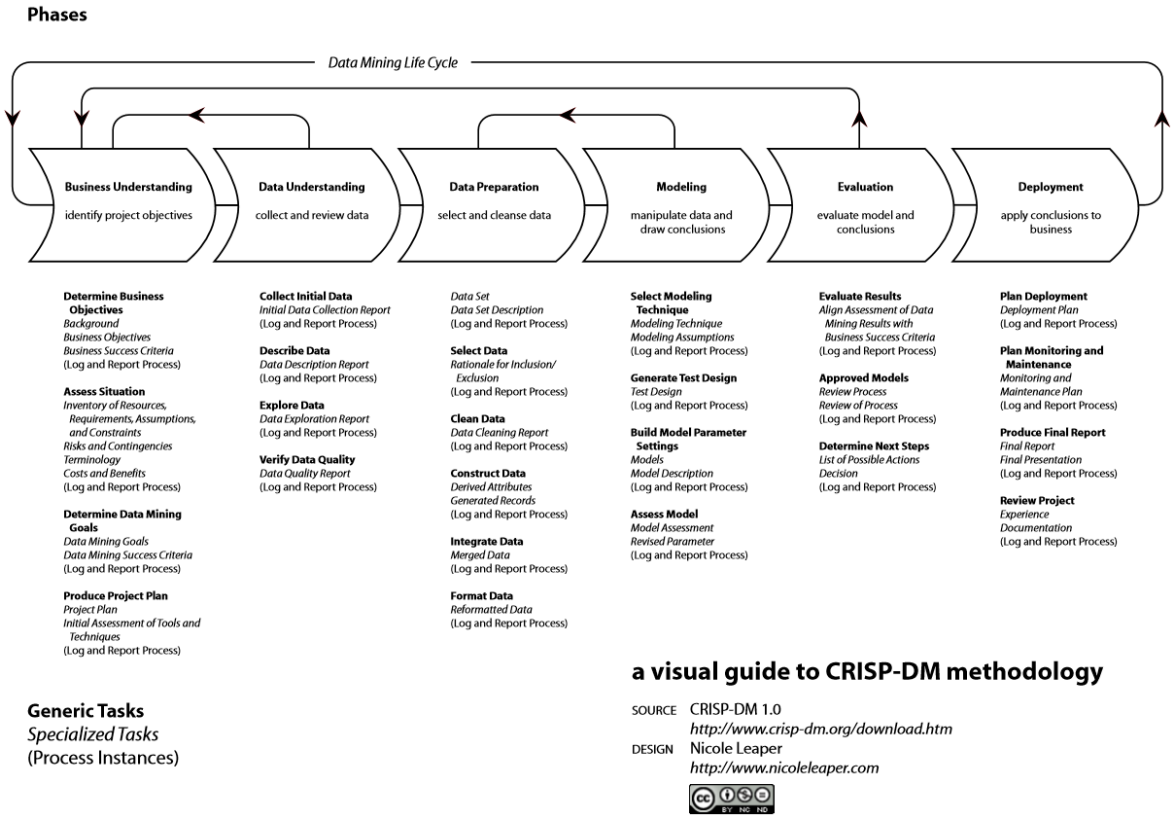


Figura 2.6: Diagrama del CRISP-DM Process

Veiem doncs en la figura 2.6 que el mètode CRISP-DM és iteratiu. Apareix també a la imatge una descripció esquematitzada de cada un dels processos a seguir durant el flux de CRISP-DM, tot i això, es realitza a continuació una explicació breu de cada una de les fases rellevants pel nostre projecte, és a dir, les tres primeres.

- *Comprenssió del negoci.* Aquesta fase inicial s'enfoca en la comprensió dels objectius de projecte. Després es converteix aquest coneixement de les dades en la definició d'un problema de mineria de dades i en un pla preliminar dissenyat per assolir els objectius.

- *Estudi i comprensió de les dades.* La fase d'entesa de dades comença amb la selecció i extracció de dades inicial i continua amb les activitats que permeten familiaritzar-se amb les dades, identificar els problemes de qualitat, descobrir coneixement preliminar sobre les dades, i / o descobrir subconjunts interessants per formar hipòtesis pel que fa a la informació oculta.

- *Anàlisi de les dades, selecció de característiques i tractament de les mateixes* La fase de preparació de dades cobreix totes les activitats necessàries per construir el conjunt final de dades (les dades que s'utilitzaran en les eines de modelatge) a partir de les dades en brut inicials. Les tasques inclouen la selecció de registres i atributs, així com la transformació i la neteja de dades per a les fer-les aplicables a les eines de modelatge.

Observem doncs què el CRISP-DM, ens proposa un sistema de flux semblant al proposat per KDD però que, al ser iteratiu, ens dóna un valor afegit pel que fa a flexibilitat. A més a més, afegeix fases dedicades

a la compressió i selecció de les dades i també a la gestió i planificació del pla de projecte a realitzar. Per tot això, creiem que ens basarem en el sistema CRISP-DM pel que fa a la generació del flux de processos, ja que, s'ajusta molt més a les necessitats del projecte plantejat. [14]

Concloem aquest apartat determinant com a rellevant el DMP CRISP-DM de cara a la realització del projecte. Es pretén, per tant, plantejar un sistema que pugui funcionar amb un flux de processos CRISP-DM o amb un flux molt semblant amb variacions minoritàries donades per l'especificitat del cas en concret de treball. A més a més de tots els factors explicats per la selecció d'aquest DMP, també s'ha tingut en compte la rellevància que té el sistema CRISP-DM, que com el seu propi nom indica, és un estàndard dins dels sectors del data mining.

2.3.3 Software existent

Finalment parlarem dels diferents softwares que es troben al mercat que permeten automatitzar els diferents processos plantejats per CRISP-DM. Donem doncs les valoracions dels dos softwares que més s'ajusten a les nostres necessitats dins de l'informe emes per The Forrester Wave™ [15] d'on s'ha seleccionat Alteryx i RapidMiner. Cal remarcar que s'han descartat softwares basats en llenguatges, ja que, es pretén realitzar una automatització user friendly i no pas donar un llenguatge imperatiu per a la interacció amb la base de dades. Tot i això, no es descarta l'ús d'aquest tipus d'eines amb un front-end que permeti la realització de les tasques sense coneixements en programació. Un cop aclarit, comencem exposant el cas de RapidMiner.

RapidMiner és una eina que ens permet aplicar diferents accions sobre els datasets amb una GUI interactiva, resulta molt fàcil d'entendre'n el funcionament i existeix documentació exhaustiva de cada mòdul que implementa. Aquest software permet generar un flux d'execució en el que és dona la possibilitat de tractar les dades, aplicar el modelatge, realitzar cross-validations, etc. Resulta una eina interessant pel fet de la facilitat que ofereix a l'usuari, que tan sols comunicant caixetes (mòduls amb inputs i outputs) pot aconseguir aplicar tot un sistema de l'estil KDD (ignorant la primera fase d'extracció de dades) i anar-lo executant variant els valors dels paràmetres. El principal punt a favor però és la corba d'aprenentatge d'aquest, resultant molt fàcil d'aprendre respecte d'altres softwares. Una de les inconvenients què té és que hi ha limitacions, és a dir, en basar-se en mòduls implementats per casos generalistes sovint no són plenament funcionals per casos específics. [15]

Resulta doncs ser una possibilitat viable, l'ús de RapidMiner per a la realització del tractament de dades i com a possible solució a l'aplicació del modelatge. Passem ara a la segona opció a tractar, Alteryx, presentant la definició que ells mateixos ofereixen.

> **Citació 2** ***Alteryx:*** *"Without the right data preparation tools, analysts struggle with the data preparation work need for analysis. Various data types/sources and disparate point solutions can further compound the challenges. Alteryx enables self-service data analytics through a complimentary set of data preparation tools that speeds up the data preparation process to deliver the right dataset to drive downstream process or improve analytic modeling and reporting. Alteryx's data preparation tools allow analysts to:*
>
> - *Connect to, prepare and combine data from various sources with drag-and-drop tools*
> - *Improve data integrity and quality with tools like fuzzy matching and parsing for advanced cleansing*
> - *Create a repeatable workflow design that speeds up the data delivery process.*
>
> *[16]*

Observem aquesta vegada que aquest software realitza una funció molt semblant a l'oferta per RapidMiner, donant un sistema amb GUI per al tractament i l'estudi de les dades. A més a més, Alteryx dona possibilitats més exhaustives de preprocessing i encreuament de dades. Tot i això Alteryx és un software que no dóna un rendiment del mateix nivell que RapidMiner.

Un altre punt que s'ha derivat de la investigació d'aquestes dues eines és la llibreria KNIME de python. S'ha descobert aquesta llibreria que sembla aportar un gran nombre d'opcions i que tot i anar directament relacionada amb codi, sembla ser prou clara perquè una persona amb coneixements molt bàsics de programació l'apliqui. Sembla que dóna un rendiment molt superior a les altres opcions i que, al no estar basat en mòduls genèrics sinó en mòduls programables, ofereix un abast molt més gran que les altres eines presentades. [17]

Concloem doncs en aquest apartat que després de realitzar l'estudi dels softwares en ús s'haurà de contemplar la possibilitat d'usar eines com les presentades en cas que el seu rendiment pugui arribar a ser superior al de les usades actualment. Per tant, possiblement, s'acabarà realitzant una anàlisi exhaustiva d'aquestes eines per veure si són aplicables al nostre projecte o si poden solucionar alguna de les diferents parts del tractament de les dades amb major eficiència que les eines existents actuals.

3

Definició i abast del projecte

Una vegada estudiat el context en el qual ens trobem i situats en el problema existent passem a definir el projecte a realitzar. Per començar és dona una descripció que intenta donar una idea general del projecte a realitzar.

El projecte a realitzar és el disseny d'un sistema de base de dades i un sistema de migracions de dades que permeti emmagatzemar i unificar totes les dades que es reben al departament de 'Data analytics' de l'empresa IKI MEDIA COMMUNICATIONS, S.L. per a facilitar-ne l'ús i assegurar-ne la integritat. Les dades rebudes provenen de diferents serveis i fonts, fet que fa que sovint puguin mantenir una mateixa informació en diferents formats, que gairebé sempre no siguin referents a un mateix àmbit temporal, que freqüentment incorporin errors de coherència i també que existeixin casos en els quals ens trobem dades que s'han de tractar prèviament abans de ser útils a causa d'errors en les mateixes o falta de completesa d'aquestes. Tot això acaba fent que el fet de creuar aquestes dades no sigui trivial i el procés de tractament sigui llarg i laboriós. El projecte per tant, té com a objectiu, mitjançant l'estudi dels processos actuals, trobar una proposta de disseny que permeti crear una base de dades (que per requisit de l'empresa haurà de ser relacional) que mantingui totes les dades que es reben mantenint la coherència i realitzar una reestructuració dels processos per tal de mantenir-la i actualitzar-la diàriament amb la informació rebuda. Eventualment en cas de ser necessari també es facilitarà informació per tal de facilitar la migració inicial de dades actual. El projecte doncs, espera basar-se en un procés semblant al següent:

1. Estudiar el funcionament tant en l'àmbit de processos com pel que fa al software de l'obtenció i el tractament de les dades. (Estudi de context).
2. Realitzar una anàlisi complet dels requisits necessaris de la proposta donat els resultats de l'estudi previ. (Anàlisi de requisits).
3. Disseny de la proposta (Disseny del software i disseny de l'arquitectura).
4. Informe del flux de processos per a l'ús de la nova base de dades i realització de la documentació del sistema (Documentació i demostració).

Val a dir que tot i trobar-se fora de les intencions inicials, en cas de tenir temps es realitzarà la implementació i testeig d'un prototipus que permeti realitzar una primera prova de concepte en local del sistema plantejat.

3.1 Abast del projecte

Per tant, l'abast d'aquest projecte serà el disseny d'un sistema de base de dades que permeti extreure, tractar i emmagatzemar les dades de manera neta i completa assegurant que posteriorment és podran usar per a la realització d'estudis analítics. Per a la realització d'aquest disseny serà necessari passar per diferents stages de les metodologies típiques de l'enginyeria del software com poden ser, l'estudi de context, incloent-hi anàlisi de stakeholders exhaustiu, anàlisi de hardware existent, anàlisi de software en ús, etc. L'anàlisi de requisits, el disseny d'un sistema (tant software com hardware) que permeti assolir els requisits trobats i proposar un nou sistema de funcionament que mitjançant l'ús del sistema dissenyat millor que el sistema de funcionament actual.

Inicialment per tant, queda fora del scope la implementació i testeig d'un prototipus com a prova de concepte, però si les fites temporals ho permeten, es realitzarà un petit prototipus que doni l'oportunitat de veure el sistema en funcionament. D'altres coses que podrien resultar dubtoses de trobar-se o no trobar-se dins del scope del projecte però que inicialment no hi són contemplades poden ser: justificació dels datasources que s'usen, automatització de les anàlisis bàsiques de les dades, justificació dels softwares en ús de l'empresa.

3.2 Objectius

Un cop realitzada la definició de l'abast del projecte passem a definir els objectius que es pretenen complir amb la realització d'aquest treball de final de grau i delimitar encara més l'abast del projecte. Per tal de fer-ho s'ha decidit usar objectius S.M.A.R.T.(Specific, Measurable, Achievable, Relevant, Time bound) [18] que ens permetran especificar objectius clars, assolibles i significatius. En el cas d'estudi el T (Time-Bound) sempre serà l'entrega del treball de final de grau, per tant, ens estalviarem d'escriure-ho en cada objectiu deixant-ho citat aquí. Enumerem doncs els objectius del projecte a realitzar:

1. Obtenir el conjunt de requisits del sistema a dissenyar i demostrar el compliment d'aquests en el disseny final. *Measure:* Existeix una llista de requisits i una justificació del compliment de cada un en el sistema dissenyat.

2. Realitzar un informe de la situació actual que permeti detectar possibles errors de mètode i ajustar el sistema a les necessitats específiques. *Measure:* Existeix un estudi complet de procés a partir del qual es justifiquen decisions del sistema dissenyat.

3. Dissenyar un sistema que permeti emmagatzemar les dades extretes en un format estàndard. *Measure:* Permet emmagatzemar tots els tipus de dades existents actualment.

4. Dissenyar un sistema que permeti extreure i emmagatzemar les dades d'estudi automàticament de manera diària/setmanal/mensual [depenent de la font]. *Measure:* S'ha dissenyat un sistema que permet recollir les dades de totes les fonts automàticament.

5. Proposar un flux de processos que permeti realitzar la feina actual suposant l'ús del sistema dissenyat. *Measure:* S'ha donat una proposta de nou flux de processos que permet obtenir els mateixos resultats que actualment usant el nou sistema plantejat.

3.3 Restriccions inicials

Un altre punt a tenir en compte en aquest projecte és que en formar part d'una empresa amb uns recursos limitats i un funcionament intern determinat existeixen restriccions que haurà de complir el projecte que també haurem de tenir en compte. Per tant, les enumerem a continuació:

1. El sistema ha de ser integrable en els softwares en ús (R, Excel, Tableau, etc.)
2. El sistema s'ha de poder integrar en futurs softwares de l'empresa, és a dir, integrable en Java i VB0.
3. El sistema no pot contractar servidors ni estar basats arquitectures físiques no aplicables amb els servidors ja existents.
4. S'ha d'usar un sistema de bases de dades relacional.

3.4 Principals inhibidors

Els principals inhibidors que podem arribar a trobar-nos en aquest projecte són la mala comunicació i el conflicte d'objectius.

El primer el justifiquem, ja que, per a la realització d'aquest projecte es necessita una comunicació molt directe i continuada amb els treballadors del departament, especialment de cara a l'estudi del context. Si bé, podem confirmar que la relació amb ells és molt correcte i què compartim espais, existeix la possibilitat que de la gran quantitat de feina que es rep no deixi temps suficient per a realitzar la quantitat de reunions necessàries per a la comprensió completa del context i la proposició d'un bon sistema que encaixi amb tots els stakeholders interns.

Pel que fa al segon també va relacionat amb la càrrega de feina dels treballadors del departament, si bé és cert que tots estan d'acord amb la necessitat d'una millora tecnològica que redueixi la quantitat de feina manual, incrementi l'eficiència i redueixi l'error humà. Podria passar que, en algun moment de molta càrrega de treball, els seus objectius es modifiquessin i fossin usar el meu temps per a treure la feina del moment en comptes de ser per a dissenyar un sistema.

La solució que trobem en ambdós casos recau en mantenir una comunicació el més constant possible amb tots els components de l'equip per adaptar-nos al màxim a les necessitats individuals i tenir disposició a mantenir comunicació fora de l'horari laboral per tal d'assegurar l'èxit del projecte. Aquesta proposició ha estat parlada amb tots els components del departament i s'han acceptat aquestes condicions.

4

Metodologia

Per a la realització d'aquest projecte s'ha decidit usar una metodologia àgil Scrum[19]. La decisió no ha estat arbitrària sinó que s'ha decidit usar aquest tipus de gestió a causa del fet que com s'ha explicat (2.2.4 - Estudi context inicial: Stakeholders: Estadístics) existeix una visió entre els estadístics de resistència al canvi a causa de males experiències, així doncs el mètode àgil ens permetrà donar petits increments de feina per ensenyar, que facilitaran l'adaptació a la idea del nou sistema. Aquest sistema de funcionament també ens servirà per poder realitzar i adaptar-nos a canvis en meitat del projecte, ja que, les metodologies àgils permeten iterar sobre els documents per actualitzar-los en dependència a les necessitats de cada moment.

Expliquem breument el funcionament de la metodologia àgil Scrum. Per realitzar-ho s'usarà la figura 4.1, que ens permetrà donar una explicació genèrica dels mètodes àgils.

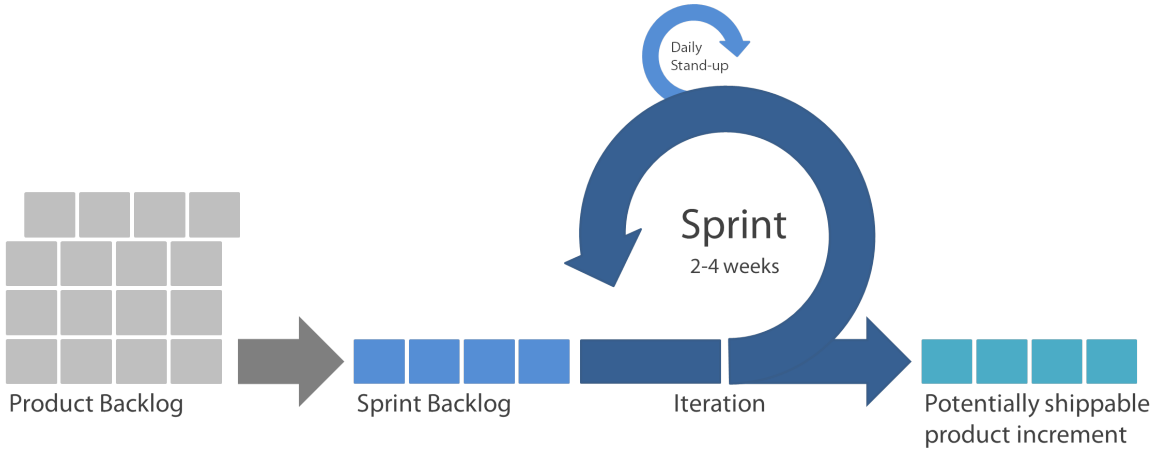


Figura 4.1: Flux d'actuació Scrum

Veiem a la imatge que tot comença amb un element anomenat Product Backlog. El product backlog és un conjunt d'històries d'usuari (les històries d'usuari són els requisits dels projectes en cascada o clàssics). Aquest product backlog representa doncs, tots els requeriments que s'han de complir per donar per finalitzat el projecte. El product backlog es genera en una primera fase del projecte anomenada 'inception' què és on es decideixen tots els requeriments i costos temporals d'aquests. Un cop passada aquesta primera fase, els projectes Scrum iteren en forma de sprint, és a dir, es realitzen petits projectes d'entre 1 i 4 setmanes en els que s'implementen o es porten a terme un subgrup de les històries d'usuari plantejades al product backlog, generant al final del sprint un subproducte final amb les tasques realitzades dins del

sprint acabades i preparades per donar a client. Per tant a cada sprint s'afegeix valor al producte final fins que s'acaben de realitzar totes les històries d'usuari existents al product backlog. Existeixen 3 reunions necessàries per a tot projecte Scrum:

- Sprint planning meeting. Reunió en la qual es decideix quines històries entren al backlog del sprint.
- Sprint review meeting. Reunió en la qual es realitza el tancament d'un sprint i es valora el funcionament d'aquest.
- Stand up meeting. Reunió diària en la que els components de l'equip informen del seu estat i es solucionen els problemes d'execució diaris.

Passem ara al plantejament específic d'aquest projecte. La durada dels sprint d'aquest projecte és d'una setmana. Durant el sprint s'esperen com a mínim dues 'stand-up meeting' i les reunions de 'sprint planning' i 'sprint review' en format reduït, recordem que existirà una planificació inicial de totes les tasques forçada per l'assignatura de GEP. També s'espera un sprint d'una setmana i mitja o dues setmanes (depenent de la càrrega de feina) dedicat a la incepció en el que es determinaran totes les tasques i el seu cost en temps mitjançant un 'poker planning', fet que ens permetrà portar una gestió temporal del projecte que s'explicarà en el següent punt. A més a més, es pretén fer entendre a l'equip el concepte de feature per substituir el de requeriment i tractar completament la gestió de forma àgil.

4.1 Seguiment evolució

Prèviament s'ha plantejat la realització d'un 'poker planning' que ens permetrà assignar una durada a les tasques que també s'usarà per a la realització del seguiment. Inicialment al final del primer sprint d'incepció es decidirà el nombre total de sprints a realitzar durant el primer sprint review i a partir d'aquí es dividirà el total d'hores de càrrega aconseguides durant el 'poker planning' per saber la suposada càrrega mitjana que cada sprint hauria de tenir. Posteriorment a cada sprint planning es decidiran les tasques que entren dins d'aquell sprint que hauran de tenir una càrrega total semblant a la mitja calculada anteriorment. Un cop assignades les tasques d'un sprint es farà un diagrama de gantt que ens permetrà gestionar les tasques dins del límit temporal del sprint.

Pel que fa al seguiment d'avenç del projecte es realitzaran dos artefactes que permetran aquesta avaluació. El primer seran Burn-down charts, basats en les hores de càrrega assignades a cada tasca, que es realitzaran tant per sprint com pel que fa a product backlog. A més a més també es farà un diagrama de gantt que mostrarà l'avenç real de les tasques, és a dir, que permetrà una comparació amb el realitzat a inici de sprint, veient si el projecte avança al ritme esperat.

5

Planificació del projecte

La duració estimada del projecte és d'aproximadament 16 setmanes. El projecte de final de grau s'inicia el dia 26 de febrer de 2018 i s'ha d'entregar abans del 18 de juny de 2018. En aquest cas específic se suposa una càrrega setmanal de 30 hores, 6 hores de càrrega diària, en horari laboral més 12 hores extra setmanals fora de l'horari laboral dedicades a GEP.

Amb l'objectiu de tenir marge d'error i de poder afrontar els problemes que sorgeixin sense incrementar el nombre d'hores a realitzar diàriament, ens posem com a data límit de finalització a la nostra planificació el dia 1 de juny de 2018, fet que, en cas de trobar-nos davant un problema què alenteixi la realització del projecte ens permetrà seguir dins del marge d'entrega real fixat per la FIB. A més, en cas d'entrar dins del termini què ens establim (26/02 a 01/06), ens oferirà un marge de temps per a la revisió i la implementació d'un prototipus del disseny plantejat afegint un valor extra al projecte.

També val la pena recordar que funcionem amb un sistema àgil i que, per tant, aquesta planificació pot ser modificada degut a noves necessitats. Per tant aquesta planificació resulta ser un document iterable sobre el qual s'ha de treballar en cada un dels sprints.

5.1 Definició de les tasques

En aquest punt es definiran totes les tasques identificades a realitzar durant aquest procés. Recordem però què aquestes poden tenir alteracions o que en poden aparèixer de noves a causa de la metodologia usada. S'han classificat les tasques en tres grups diferents que pretenen separar el projecte en les tres fases del disseny de software: Estudi de context, Anàlisi de Requeriments i Proposta d'una solució. Afegim un punt al final de cada tasca dedicat a enumerar els recursos necessaris per a la feina, facilitant així la futura comprensió del pressupost, no s'inclouen els recursos de hardware, ja que són necessaris per a totes les tasques.

5.1.1 Estudi de context

Aquest paquet de tasques, serà probablement el què més interacció i recursos necessitarà, doncs per a l'estudi del context, serà imprescindible comptar amb la visió de cada un dels 'stakeholders' interns i d'alguns 'stakeholders' externs. Les tasques dins d'aquest apartat són:

5.1.1.1 Stakeholders interns (A)

En aquesta tasca es pretén realitzar un estudi en profunditat dels diferents 'stakeholders' interns anomenats en l'apartat 'stakeholders'. L'objectiu d'aquesta tasca és aconseguir extreure la informació sobre

les necessitats reals de cada un dels 'stakeholders' interns per tal de poder intentar adaptar la solució al màxim a cada un d'ells. Per la realització d'aquesta tasca per tant, es necessitarà tenir entrevistes personals amb cada un dels diferents 'stakeholders' per separat i una de conjunta amb l'equip per tal d'extreure la informació pertinent als objectius de cada un amb el projecte i alhora els objectius generals, també s'espera trobar l'ordre de prioritat d'aquests objectius. Un cop realitzades aquestes reunions es tractarà d'ajuntar la informació recopilada i d'extreure tant una anàlisi 'stakeholder' per 'stakeholder' com una anàlisi general d'equip amb l'ordre de prioritat de les necessitats trobades.

Per tant, aquesta tasca és divisible en dues subtasques que són:

- *Recollida informació amb entrevistes:* Es pretén entrevistar a les dues persones dedicades a Data Science, a l'analista programador i a l'equip directiu per separat. També es busca realitzar una reunió general amb representants de cada un dels grups prèviament citats. S'han seleccionat aquests 'stakeholders', ja que, els resultats del punt (*Priorització Poder vs. Interès dels 'stakeholders'* varen situar-los com els 'stakeholders' interns més prioritaris.

- *Realització estudi de la informació aconseguida* L'objectiu d'aquest punt és, a partir la documentació de les entrevistes realitzades, obtenir una anàlisi complet d'objectius de cada 'stakeholders' amb les seves prioritats establertes i documentació dels objectius comuns entre 'stakeholders'.

Recursos necessaris: Estadístics, analista programador, equip directiu, material d'oficina.

5.1.1.2 Estudi Data Sources (B)

La tasca d'estudi dels datasources es basarà en la realització d'una anàlisi de les diferents fonts de dades que s'usen per a l'obtenció de les dades a treballar. En aquesta tasca primerament es buscarà determinar totes les fonts de dades que s'usen. Un cop determinades es realitzarà un estudi genèric de cada una de les fonts que ens permetrà saber la informació que s'obté de cada una de les opcions contractades. Posteriorment es demanarà a cada un dels proveïdors de dades el fitxer de metadades pertinent a cada font i en cas de no rebre resposta o la negació de l'enviament d'aquest fitxer, es realitzarà manualment la descripció camp a camp, per tal de generar el fitxer. Finalment i un cop obtingut aquest fitxer, es realitzarà un disseny UML font a font que permeti visualitzar una possible estructura d'emmagatzemament de les dades. Per a aquesta tasca doncs, es necessitarà tan sols la realització d'una reunió dins del departament per tal de determinar les fonts de dades usades, una reunió amb l'equip de direcció per saber les fonts de dades contractades i el contacte via e-mail amb les diferents propietàries de les fonts de dades.

Recursos necessaris: Telèfon oficina.

5.1.1.3 Estudi del software en ús i del hardware (C)

En aquest punt es busca aconseguir un petit estudi que permeti determinar quins softwares es fan servir durant el procés. L'objectiu d'aquesta tasca doncs és determinar a quins softwares s'haurà d'ajustar la proposta i veure quin és el potencial de cada un d'aquests per determinar si podrien ser substituïts per altres softwares o simplement poden deixar de ser usats. Per a la realització d'aquesta tasca es necessitarà una reunió amb l'equip en la qual es demanarà quins softwares s'usen, per què, i la necessitat d'ús (subjectiva) de cada un d'ells actualment. També es busca obtenir informació bàsica del hardware què l'empresa té. Per fer-ho es mantindrà una reunió amb l'analista programador.

Recursos necessaris: Estadístics, analista programador, equip directiu, material d'oficina.

5.1.1.4 Estudi Procés Actual (D)

Aquesta tasca pretén obtenir com a resultat un diagrama i documentació què ens permeti identificar i entendre el sistema de processos que s'esta portant a terme actualment per a la realització de la feina del departament de Data Science. Aquest estudi pot resultar llarg, ja que s'haurà de realitzar almenys un parell de reunions amb cada un dels diferents 'stakeholders' interns rellevants segons els estudis (Data Science, analista programador i equip directiu). L'objectiu de la primera reunió serà entendre i modelitzar el seu procés de treball i en la segona es presentarà el model de processos per revisar si la feina realitzada encaixa amb la feina real que cada 'stakeholder' realitza. En cas de no ajustar-se correctament, es redissenyarà i es tornarà a mostrar al 'stakeholder' iterant fins que el resultat encaixi amb la realitat. Finalment es presentarà una versió completa dels processos que passarà per un sistema de revisió basat en una reunió amb representants de cada tipus de 'stakeholders' que revisaran la correctesa del model de processos plantejat. D'aquesta última reunió s'espera sortir amb un diagrama de processos correcte i amb informació suficient per a la realització de la documentació complementària als diagrames.

Recursos necessaris: Estadístics, analista programador, material d'oficina.

5.1.2 Anàlisi de requeriments

Aquest paquet, es defineix com una sola tasca (E) en la qual es pretén, a partir de la informació recopilada durant l'estudi de context realitzat, aconseguir el conjunt de requeriments necessaris que el software a dissenyar haurà de complir per tal de ser funcional. Aquest fet ens permetrà mostrar al conjunt de 'stakeholders' el que realitzarà el sistema i el que no realitzarà (una mena de scope) per tal d'evitar futurs conflictes d'objectius (evitant així també un dels principals inhibidors), ja que, a partir d'aquest moment els requisits de la solució estaran fixats i tan sols es modificaran en cas de trobar una justificació clara de millora de cara al resultat final (recordem que la metodologia àgil permet la modificació del scope durant el transcurs del projecte). Per a fer aquesta tasca, tan sols es necessitarà realitzar una reunió interna a l'empresa en la qual es mostraran els requisits del sistema per a assegurar que els requeriments s'ajusten a tothom i que hi ha un acord entre tots els 'stakeholders' en què els requisits presentats són correctes.

Recursos necessaris: Estadístics, analista programador, equip directiu, material d'oficina.

5.1.3 Proposta de solució

Aquesta part és probablement la que menys recursos requerirà, ja que es basa en la realització d'una proposta a partir de les dades prèviament recollides i estudiades. Tan sols es requerirà una reunió amb tots els 'stakeholders' interns un cop finalitzades totes les tasques del procés per tal de presentar la proposta i defensar-la.

5.1.3.1 Selecció de softwares, hardwares i datasources (F)

L'objectiu d'aquesta tasca és el de determinar quins softwares, hardwares i datasources s'usaran com a base per al disseny de la proposta a realitzar. Per a seleccionar els softwares i hardwares s'usarà la documentació recopilada en l'estudi de softwares i en l'estat de l'art prèviament realitzats. Pel que fa als datasources se seleccionaran en dependència a l'estudi realitzat en la tasca d'estudi dels datasources. Per a la realització d'aquesta tasca doncs no es necessitarà cap recurs.

Recursos necessaris: Material oficina.

5.1.3.2 Disseny de la solució (G)

A partir de totes les dades recollides, es realitzarà un disseny de la solució que es vol proposar a l'empresa. Per a la realització d'aquesta proposta s'haurà de realitzar tant un disseny del software proposat com un disseny del hardware. El resultat d'aquesta tasca, serà bàsicament la proposta de millora oferta a l'empresa i el resultat de tot el treball, és important doncs que aquest disseny sigui basat en decisions completament justificables a partir de l'estudi previ o a partir de fonts fiables. Aquesta tasca pot requerir bibliografia per a la justificació de les decisions de disseny.

Recursos necessaris: Material oficina.

5.1.3.3 Disseny de procés (H)

Finalment un cop dissenyada una solució, haurem de realitzar la tasca de disseny de procés, en què es busca aconseguir un diagrama de procés que permeti mostrar com integrar el nou sistema en la feina i demostrar quin és el flux de treball que s'haurà de seguir per a obtenir resultats semblants o millors als prèviament oferts per l'empresa. Amb aquesta tasca es pretén aconseguir un diagrama i documentació del futur flux de treball en cas d'implementació i ús de la proposta realitzada.

Recursos necessaris: Material oficina.

5.2 Estimació temporal

Paquet	Tasca	ID	Temps (h)	Depen.
Estudi de context	Estudi stakeholders interns	A	70	-
	Estudi data sources	B	40	-
	Estudi softwares en ús i hardware	C	40	-
	Estudi procés actual	D	80	B,C
Anàlisi requisits	Anàlisi requisits	E	30	A,D
Proposta de solució	Selecció software, hardware i datasources	F	30	E
	Disseny de la solució	G	70	E
	Disseny del procés	H	50	F,G
	Presentació final	I	10	H
Marge	Implementació prototip / Temps marge	F	60	*
TOTAL			**480**	

Taula 5.1: Estimació temporal de les tasques

Veiem doncs a la Taula 5.1 què, el temps total de realització de les tasques és de 480 hores. Sabem però que a aquestes 480 hores s'hi han sumat també les dues setmanes de marge que es donen per a resoldre possibles desviacions temporals i que, per tant, la càrrega que suposaran les tasques definides seran 60 hores inferior a l'estipulada inicialment, i que posteriorment s'ompliran sigui degut a desviacions o per la implementació d'un prototipus del projecte. A més a més a tot això cal sumar la càrrega representada

per GEP, d'unes 75 hores. Aquestes 75 hores dedicades a GEP, seran distribuïdes durant les 6 primeres setmanes de projecte suposant un increment de 2.5 hores diàries de treball. S'ha acordat amb l'empresa que 1 hora de dedicació serà en horari laboral i l'altra hora i mitja es realitzarà fora d'horari laboral.

5.3 Diagrama de gantt

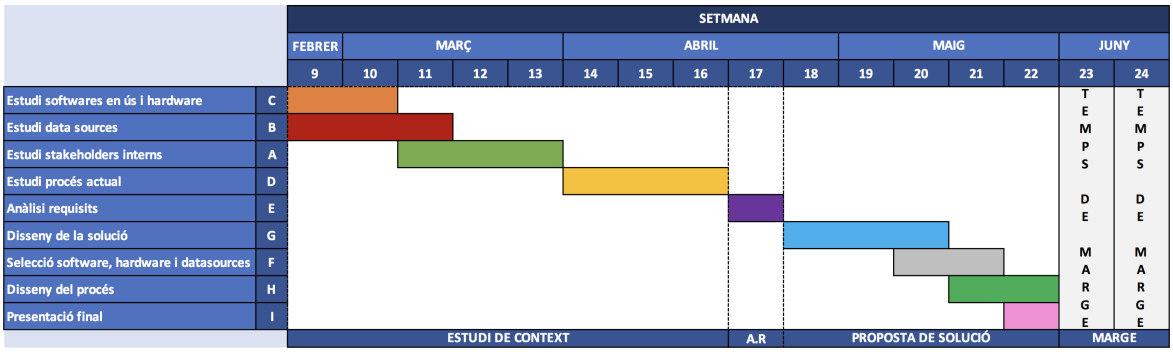


Figura 5.1: Diagrama de Gantt

S'ha realitzat aquesta organització temporal de les tasques, respectant el que s'ha plantejat anteriorment a la Taula 5.1. Algunes tasques apareixen encavalcades entre si, aquest fet no representa què es realitzin en menys hores sinó que les hores de dedicació de la setmana en la qual s'encavalquen, es dividirà entre les dues tasques per a facilitar la seva realització.

5.4 Pla d'acció

Finalment expliquem el pla d'acció que permetrà assolir la planificació realitzada. La idea principal és cenyir-nos a la planificació establerta, però és ben sabut què en gairebé tots els projectes existeixen casos que sovint alentesquen el ritme i no permeten seguir la planificació prèvia.

En el nostre cas, el punt més dèbil i el que més ens pot afectar és que, per a la realització del treball i especialment del primer paquet, és a dir, l'estudi de context, necessitem tenir moltes interaccions amb diferents persones. Fet que dificultarà l'organització temporal de les reunions, ja que, tots tenen unes agendes molt ocupades. Amb l'objectiu de solucionar-ho s'han posat les principals tasques d'interacció encavalcades, com pot ser el cas de la tasca C i B que estan temporalment sobre la mateixa setmana, ja que, en cas de no aconseguir una resposta ràpida de les fonts de dades, podrem treballar en l'estudi del hardware i del software durant el temps d'espera. Un altre mesura que s'ha pres ha estat donar un temps de marge que ens permeti, en cas d'haver d'allargar una tasca tenir fins a 60 hores de flexibilitat que ens permetran reajustar-nos. A més a més aquest fet ens dóna l'opció d'implementar un prototipus en cas de no patir desviacions temporals. Una altra solució que s'ha trobat al problema, com ja estava comentat en la descripció de la metodologia, ha estat donar flexibilitat horària per a la realització de reunions, incloent-hi hores fora de les estipulades per contracte què seran compensades en altres moments. Com a conseqüència s'espera facilitar les reunions amb els diferents stakeholders.

Per tant i basant-nos amb les dues propostes (marge d'hores i flexibilitat horària) de contingència davant dels inhibidors podem dir que asseguren que el projecte disminuirà el seu risc de patir una afectació d'aquests. Pel que fa a l'inhibidor de la diferència d'objectius, s'intenta de solucionar incrementant el nombre de reunions al màxim per tal d'assegurar cada poc temps que els objectius dels diferents stakeholder no es troben contraposats. Un altre punt què disminueix la possibilitat que aquest fet es doni, és que, a meitat de projecte es prepara una tasca d'anàlisi de requisits, a partir de la qual, els diferents stakeholders ja tindran coneixement sobre el qual realment aconseguiran amb aquest projecte i els objectius finals d'aquest.

El projecte consta de 420 hores planificades, 60 hores variables de marge que s'estan deixant per a possibles desviacions i unes 75 hores de la realització de GEP. Acumulant un total de 555 hores. Realitzant una mitja de 6 hores diàries, durant les 16 setmanes de duració del projecte, es realitzaran 460 hores en total. Veiem doncs que falten 75 hores, justament les planificades per GEP. Aquestes hores es treballaran fora de la jornada laboral (en part), fent unes 6 hores extres setmanalment entre el 26/02 i el 09/04. Concloem doncs que el projecte és realitzable amb el temps donat.

6

Estimació del cost

Un cop plantejat el pla d'acció del projecte passem a realitzar un pressupost i anàlisi de costos per a la realització d'aquest projecte. Es divideixen els costos en costos directes, indirectes i possibles incidents per al seu càlcul i cada secció conté la justificació de la pressupostació feta a cada ítem.

6.1 Costos directes

Comencem calculant el cost en recursos humans que suposarà la realització d'aquest projecte. Amb l'objectiu de fer-ho, trobem necessari dividir les hores de treball assignades a cada tasca entre els diferents treballadors, ja que, no tots tenen una mateixa base salarial. Per tant:

Paquet	Tasca	ID	Temps (h)	Estadístics	Analista Programador	Directiu	Enginyer Software	Preu tasca
Estudi de context	Estudi stakeholders interns	A	80	5	3	2	70	2.750,00 €
	Estudi data sources	B	43	2	0	1	40	1.425,00 €
	Estudi softwares en ús i hardware	C	50	4	6	0	40	1.700,00 €
	Estudi procés actual	D	96	8	4	4	80	3.500,00 €
Anàlisi requisits	Anàlisi requisits	E	30	0	0	0	30	900,00 €
Proposta de solució	Selecció software, hardware i datasources	F	30	0	0	0	30	900,00 €
	Disseny de la solució	G	70	0	0	0	70	2.100,00 €
	Disseny del procés	H	50	0	0	0	70	2.100,00 €
	Presentació final	I	16	2	2	2	10	750,00 €
Marge	Implementació protopip / Temps marge	F	60	0	0	0	60	1.800,00 €
TOTAL				21	15	9	500	17.925,00 €

Taula 6.1: Divisió temporal de les tasques

Veiem doncs que els estadístics tindran una dedicació d'unes 21 hores, l'analista programador d'unes 15 hores, l'equip directiu d'unes 9 hores i finalment l'enginyer del software tindrà una càrrega de 500 hores (Taula 6.1). Concloem doncs que gairebé tots els rols menys l'enginyer del software tenen una càrrega baixa, fet que, és justificable, ja que l'única tasca que hauran de portar a terme és la de mantenir reunions amb l'enginyer del software. Cal recordar també que tota aquesta estimació es fa sense tenir en compte les 60 hores reservades per a possibles incidents. En la taula també s'inclou el cost per cada tasca, per a calcular-ho s'han usat els preus per hora mostrats a la Taula 6.2. Passem doncs a calcular el cost en recursos humans total del projecte.

Persona	Hores	Preu brut per hora	Preu brut total
Estadístics	21	50,00 €	1.050,00 €
Analista Programador	15	50,00 €	750,00 €
Directiu	9	125,00 €	1.125,00 €
Enginyer Software	500	30,00 €	15.000,00 €
TOTAL			**17.925,00 €**

Taula 6.2: Cost dels recursos humans

El cost humà del projecte serà de 17.925,00 € (Taula 6.2) i s'ha extret multiplicant el nombre d'hores treballades per cada rol pel preu per hora del treballador. El preu per hora no és exacte sinó una aproximació per a mantenir la privacitat establerta a l'empresa.

Continuem ara a calculant el cost en hardware a utilitzar. S'ha suposat el material hardware mínim per a la realització del projecte, és a dir, l'ordinador d'empresa que s'ha facilitat al treballador i el telèfon per IP que s'usa per a la comunicació dins l'empresa.

Producte	Preu	Unitats	Vida útil (anys)	Amortització
iMac (Retina 4K, 21.5-inch, 2017)	1.505,50 €	1	6	83,64 €
Unify OpenStage 15 HFA V3	120,50 €	7	10	4,02 €
TOTAL	**1.626,00 €**	**-**	**-**	**87,66 €**

Taula 6.3: Cost del hardware

Obtenim doncs un cost en hardware de 1626,00 € amb una amortització de 87,66 € (Taula 6.3). Per a calcular aquesta amortització anual s'ha dividit el preu del producte entre els anys de vida útil i s'ha multiplicat per la durada de 4 mesos del projecte.

Seguim doncs ara amb el càlcul del cost en software que tindrà el projecte. Cal dir que aquest cost ha de ser el més baix possible, ja que, es demana no usar softwares de pagament perquè, el resultat d'aquest projecte no serà un software amb valor sinó una proposta de millora del sistema que actualment funciona i per tant, no hauria de tenir un impacte directe en els costos o beneficis de l'empresa a curt termini.

Software	Unitats	Preu menusal per unitat	Preu duració proyecte	Amortització
Microsoft Office 365	7	10,60 €	371,00 €	371,00 €
Latex	1	0,00 €	0,00 €	0,00 €
Draw.io	1	0,00 €	0,00 €	0,00 €
Atom.io	1	0,00 €	0,00 €	0,00 €
MacOs High Sierra	1	0,00 €	0,00 €	0,00 €
Github	2	0,00 €	0,00 €	0,00 €
Eclispe	1	0,00 €	0,00 €	0,00 €
TOTAL			371,00 €	371,00 €

Taula 6.4: Cost del software

Comprovem que el cost en software és molt baix com ja esperàvem pels motius explicats. El cost és de 371,00 € i té una amortització de 371,00 € mensuals (Taula 2.4). Resulta lògic que l'amortització resulti el mateix que el preu final, ja que, les llicències de pagament usades, són de pagament mensual i per tant, tenen una vida útil de 4 mesos.

Finalment doncs, unim el cost obtingut en les diferents taules i obtenim:

Secció	Cost
Hardware	1.626,00 €
Software	371,00 €
Recursos humans	19.922,00 €
TOTAL	21.919,00 €

Taula 6.5: Cost directe total

Veiem doncs que el cost directe del projecte serà de 21.919,00 € (Taula 6.5).

6.2 Costos indirectes

Un altre punt a calcular són les despeses indirectes del projecte. En aquestes despeses s'ha tingut en compte l'electricitat usada, la internet i el material d'oficina usat. Per l'electricitat s'ha suposat un consum de 120W per ordinador i 75W per led d'il·luminació del departament (5). Pel que fa al material d'oficina s'ha demanat una aproximació del cost en material d'oficina i despeses similars per treball al departament de comptabilitat. A més a més s'han sumat les amortitzacions calculades prèviament com a una despesa indirecte del projecte.

Producte	Preu	Unitats	Cost
Electricitat	0,13987 €/kWh	217,80	30,46 €
Internet	22,23 €/mes	5	111,15 €
Material oficina	20 €/mes	5	100,00 €
Amortitzacions	-	-	458,66 €
TOTAL			**700,27 €**

Taula 6.6: Cost indirecte total

El cost indirecte total és de 700,27€ com podem veure a la (Taula 6.6).

6.3 Imprevistos

Finalment hem de calcular els costos per imprevist que podrien afectar el nostre projecte. Recordem que un dels nostres principals inhibidors és el fet de no trobar coincidències en horaris per a la realització de les reunions, per tant, aquest també serà un punt a tenir en compte entre els possibles incidents. Presentem doncs la següent taula de pressupostació d'incidents:

Causa	Solució	Probabilitat	Impacte Cost	Cost
No coincidir reunions	Hores extres per a reunió	30%	300,00 €	90,00 €
TOTAL			**300,00 €**	**90,00 €**

Taula 6.7: Cost dels incidents total

Podem observar que hi ha el problema de no poder coincidir en reunions. Ja es va parlar amb anterioritat que la solució seria afegir la possibilitat de fer-ho fora d'hores de treball, suposant doncs un sobrecost per l'empresa. S'ha calculat que existeix una possibilitat del 30 per cent que succeeixi, i que suposa un cost de 300 euros en total. Aquests dos nombres són basats en el fet que existeixen 3 rols dins l'empresa i la probabilitat de què en una reunió almenys un no pugui és molt alta (d'aquí el 30 per cent) i de què el cost mitjà per reunió fora d'horari serà de 100 euros i existeixen 6 tasques que requereixen interacció, suposant que a la meitat d'aquestes tasques succeeix un problema d'horaris ens surt el sobrecost de 300,00 € (Taula 6.7). Altres solucions als inhibidors han estat plantejades en altres punts del treball i tractats en altres punts de la pressupostació, com ara la menció feta al temps de marge.

6.4 Cost Total

Concloem obtenint el cost total. Per fer-ho, unim tots els costos obtinguts en els punts anteriors i un 5 per cent de contingència. S'ha decidit afegir tan sols un 5 per cent perquè gairebé tots els problemes amb què ens podem trobar tenen la possibilitat de ser tractats amb el temps de marge.

Concepte	Cost
Cost directe	21.919,00 €
Cost indirecte	700,27 €
Imprevist	90,00 €
Subtotal	**22.709,27 €**
Contingencia	5%
Total	**23.844,73 €**

Taula 6.8: Cost total

Per tant, el projecte tindrà un cost total de 23.844,73 € (Taula 6.8).

6.5 Control de pressupost

Amb l'objectiu de controlar les possibles desviacions de pressupost i disminuir-les al màxim s'ha decidit realitzar un sistema de control que permeti veure la desviació en hores que s'està patint durant la realització del projecte. Per tal de fer-ho s'ha decidit que a final de cada tasca es comprovarà el temps real que la tasca ha implicat i es calcularà el següent valor:

$$\text{Desviació de temporal tasca (h)} = TE - TR \quad \text{on} \begin{cases} TE\text{: Temps estimat.} \\ TR\text{: Temps real.} \end{cases}$$

El valor obtingut s'acumularà sobre una variable (Total desviació temporal) per a totes les tasques i s'estudiarà aquesta variable. Apareixen 3 tipus d'actuació en dependència d'aquesta variable:

- *Total desviació temporal major que 20:* En aquest cas, si encara queden hores de marge, la planificació de tasques es modificarà movent totes les tasques una setmana enrera i s'afegirà una tasca per a la realització de la feina pendent a la setmana en curs o la vinent. En cas d'haver exhaurit les hores de marge, aquella setmana, s'afegiran hores extra de càrrega a l'enginyer del software per a la realització de les tasques pendents i les planificades.

- *Total desviació temporal en [-20,20]:* En aquest cas, tot funciona aproximadament segons el planejat, i per tant, no hi haurà cap canvi de planificació.

- *Total desviació temporal menor que -20:* Aquest cas es donarà si ens trobem per davant de la planificació. En aquesta situació s'avançaran totes les tasques una setmana i s'afegiran 30 hores de marge, per a futurs inconvenients.

Amb això es pretén, no superar el temps de realització del projecte assegurant doncs que l'únic valor fluctuable del pressupost sigui el màxim d'estable possible. No hi ha contemplacions respecte les hores extres degudes a reunions, s'inclouen en els costos de contingències i s'ha parlat amb direcció decidint que es tractaran com un cost inherent al projecte, ja que, a part del tractament ja planificat pel problema, és un fet sobre el qual no es pot tenir un control.

7

Sostenibilitat : Fita inicial

Passem ara a omplir la taula d'anàlisi de sostenibilitat plantejada per a la realització de GEP. S'ha decidit retirar la columna de riscos per a facilitar-ne la visualització. En cas de trobar algun risc, s'afegirà un cop exposada la taula.

	PPP	Vida Útil
Ambiental	El projecte és una proposta de solució i per tant, no té un impacte ambiental immediat. Però es té en compte aquest punt de cara a la futura proposta resultat d'aquest TFG. Es pretén presentar una proposta què disminueixi l'impacte ambiental actual. Per a fer-ho, com ja ha estat comentat, és requeriment no incrementar el hardware existent (reusar) de cara a la proposta i és objectiu reduir el nombre de màquines necessàries per a la realització de la tasca.	Recordem que actualment es consumeix emmagatzemant les dades en diferents bases de dades i es consumeix tractant-les per tal de ajuntar-les. Per tant, la proposta a realitzar per si sola hauria de millorar aquest fet, doncs ha de ser una única base de dades centralitzada què evitarà tant tenir més d'una base de dades com el tractament de les dades per ajuntar-les. Reduint conseqüentment el consum d'energia.
Econòmica	Després de realitzar l'estudi de costos del projecte podem concloure què és viable, ja que, tot i poder semblar una inversió molt elevada d'inici. La reducció de costos que pot suposar en el futur i l'increment en el valor del servei en cas de realitzar-se recuperen la inversió gairebé de manera immediata. El fet de poder oferir informes de dades més freqüents hauria de facilitar l'entrada de nous clients i hauria de reduir considerablement el cost de la realització dels mateixos.	Actualment el procés d'ajuntat i emmagatzemament de les dades s'està realitzant de manera manual, suposant moltes hores de recursos humans fent una tasca molt monòtona. A més s'estan mantenint varis servidors amb les dades per separat . La meva proposta ha de reduir el cost en servidors de dades i sobretot ha de reduir les hores de recursos humans, ja que, ha de permetre l'encreuament de les dades automàticament. A més ha de reduir el factor d'error humà i per tant, com a conseqüència reduir el risc actual, disminuint també el temps de realització que sol ser incrementat per aquest factor.
Social	A nivell personal aquest projecte em permetrà realitzar el meu primer disseny complet de sistema real i em donarà possibilitats de gestionar el projecte de implementació de la proposta en cas de ser acceptada. A més a més m'aportarà molta experiència en l'àmbit de gestió de bases de dades i gestió de processos. També espero rebre indirectament formació bàsica en el sector de la publicitat, sobre data mining i data science.	Actualment els treballadors realitzen aquesta tasca a mà. Tan sols pel fet de proposar automatitzar-ho ja és millora qualitativament la vida de les persones encarregades d'aquesta tasca que en cas de tirar-se endavant la proposat es podran dedicar a realitzar tasques de més interès . A més a més, de cara al client, la automatització és veurà reflectida en un increment en la freqüència d'entrega d'informes millorant el servei, fets que justifiquen l'existència del projecte.

Taula 7.1: Taula Sostenibilitat (Fita Inicial)

7.1 Possibles riscos

- *Dimensió social:* un conjunt de treballadors de l'empresa poden passar a no tenir tasques a fer i com a conseqüencia quedar-se sense feina.
- *Dimensió social:* facilitar l'anàlisi de dades pot portar a fomentar l'estudi dels targets en sectors que actualment usen minoritariament el datascience facilitant dades als encarregats marketing que poden no usar èticament.
- *Dimensió ambiental:* facilitar l'anàlisi de dades pot portar a incrementar el nombre de dades a guardar i incrementar conseqüentment el consum.

8

Estudi de Context: Anàlisi de les fonts de dades

En aquest apartat es pretén realitzar una anàlisi exhaustiva de les diferents fonts de dades utilitzades. Es començarà realitzant una introducció que permeti entendre els dos tipus de font de dades que es tractaran; seguidament, es plantejarà el sistema d'avaluació de qualitat i, finalment, es procedirà a realitzar l'anàlisi de cada un dels datasources mitjançant l'avaluació de la qualitat, la presentació de les metadades i el plantejament d'un disseny UML que permeti donar una estructura de classes possible.

8.1 Tipus de fonts: Globals vs. Específiques

S'han trobat dos tipus de fonts diferenciables segons la procedència; les primeres són les que provenen de serveis contractats externament. Aquestes, tenen un seguit de característiques en comú com ara que són molt completes, mantenen un format semblant i són accessibles i mantingudes. Anomenarem aquest tipus de fonts amb el terme "GLOBALS", ja que, en podem extreure dades usables per a tots els clients. L'altre tipus de fonts que s'han trobat són aquelles específiques que arriben dels clients i de fonts gratuïtes com podria ser, per exemple, el meteocat. Aquestes dades tenen en comú que no solen ser puntuals, que tenen canvis sobtats de format i que no són plenament mantingudes ni accessibles. Així doncs, aquest segon tipus de fonts s'han anomenat "ESPECÍFIQUES", ja que, no són usables per tots els clients sinó que són diferents per a cada un d'ells. Presentem la següent figura amb la classificació de les fonts de dades a treballar en aquests dos tipus (Figura 8.1).

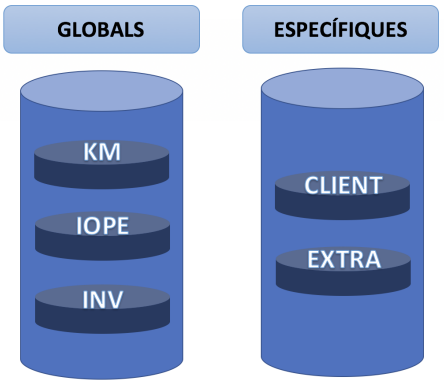


Figura 8.1: Classificació segons el tipus de les fonts de dades

A la figura 8.1 podem observar que tractarem amb 3 fonts de dades estàtiques de tipus global (Instar Analytics (KM), Infoadex (INV) i Infotrack (IOPE)) i 2 fonts dinàmiques de dades específiques: una serà la informació aportada pel client (i per tant existiran *n* fonts de dades on *n* és el nombre de clients.) i l'altra serà el conjunt de dates extra agafades per a l'estudi específic d'alguns estudis com podrien ser, per exemple, les temperatures o les precipitacions. En aquest punt veiem que haurem de tractar els dos tipus de fonts de manera diferent, ja que una roman gairebé estàtica independentment del client i de l'estudi mentre que les altres són canviants depenent del client, de l'estudi i del moment. Aquest fet ens aporta una informació molt important, ja que probablement la millor solució passarà per trobar un sistema d'emmagatzemament de les fonts estàtiques globals i, posteriorment, l'adaptació de la informació específica al sistema dissenyat. Això ens permetrà donar flexibilitat de canvi a les fonts amb més dinamisme i més estabilitat a les fonts menys canviants; és a dir, assegurant l'estructura estable de les dades globals i donant llibertat de canvi a les dades específiques.

8.2 Qualitat de les dades

Per fer l'estudi de la qualitat d'aquestes fonts s'usarà l'índex explicat prèviament a l'estat de l'art, és a dir, el CBR (Case-based Reasoning). La decisió d'usar aquest sistema d'avaluació de la qualitat no és arbitrària, sinó que s'ajusta a l'estudi a realitzar. Cal destacar que és molt important tenir una visió general de la qualitat de les dades (índex general) però, especialment en el cas d'estudi tractat, també resulta indispensable poder detectar certs problemes a nivell més específic de les dades en qüestió (fet que CBR ens permetrà a partir de l'avaluació dels subíndexs necessaris per fer el càlcul de l'índex general). De cara a poder usar aquest índex, s'han de definir quins casos base o subíndexs s'usaran per a l'estudi. En aquest punt, val la pena recordar que durant l'estat de l'art es van definir un conjunt de criteris que es podien tenir en compte com a subíndexs a utilitzar. A més, s'ha decidit que el pes total complirà la següent restricció (Figura 8.2) per a facilitar-ne la comprensió, ja que, resulta més fàcil entendre un concepte quan se'n pot parlar directament en termes de tant per cent que quan existeix una conversió intermèdia a realitzar.

$$\sum_{i=1}^{i=n} Wi = 100$$

Figura 8.2: Restricció del CBR

Un cop presentada l'única condició, passem a la selecció dels subíndexs presentats anteriorment (al punt de l'estat de l'art). Per a fer-ho, es presentaran per separat cada un dels criteris i es plantejaran les raons per les quals s'ha decidit incloure o excloure el subíndex en el cas d'estudi. A més a més, en cas de ser un subíndex seleccionat, també s'esmentarà el sistema de càlcul del valor de l'índex. Finalment, un cop acabat aquest pas, es definirà una taula a manera de resum on s'hi veuran representats els subíndexs seleccionats amb el pes assignat a cadascun d'ells. Comencem doncs amb l'anàlisi individual de cada un dels subíndexs:

- *Accessabilitat.* S'ha decidit descartar aquest criteri. Aquesta decisió s'ha degut a que cal recordar que les principals fonts de dades usades són sistemes externs contractats que asseguren l'accessibilitat de les dades però que també existeixen dades provinents del client, les quals no l'asseguren i no són descartables. Per tant, s'ha negat l'ús d'aquest subíndex ja que, aquest generaria una desviació a la baixa en els casos de les dades rebudes directament del client i una desviació a l'alça

en el cas de les fonts de dades contractades, donant una alteració en l'índex final que no seria representativa de la qualitat real pel cas d'estudi en qüestió.

- *Quantitat*. S'ha decidit seleccionar aquest criteri. Aquest fet és degut a que, de cara a l'estudi posterior de les dades, resulta necessari que la quantitat de les dades sigui ajustada a l'anàlisi; per exemple, per a la realització d'una anàlisi de l'impacte diari serà necessària una quantitat de dades major que per a la realització d'un estudi a nivell mensual. Així doncs, veiem clarament que quant més desagregades es trobin les dades més valor tindran. El sistema de càlcul d'aquest criteri es basarà en les freqüències temporals de les dades: avaluant les dades anuals amb la pitjor puntuació i les dades a nivell de segon avaluades amb la millor puntuació.

- *Credibilitat i Reputació.* S'ha decidit seleccionar aquests criteris com a agrupació. La selecció d'aquests criteris ve determinada per la importància de la fiabilitat de les dades usades; cal recordar que en cas d'usar dades errònies el resultat de totes les anàlisis estarien allunyades de la realitat i que, per tant, la pressa de descicions de cara a la inversió dels clients no estaria assegurada. Resulta de vital importància doncs, que les dades siguin creibles i que representin la realitat amb el màxim d'ajust si no és vol danyar la imatge de l'empresa. El sistema de càlcul d'aquest subíndex es basarà en la realització d'una petita investigació de reviews generals entrades a internet pels casos dels serveis de dades contractats i, pel cas de les dades provinents de clients, seran basats en les experiències prèvies amb les dades rebudes.

- *Completesa.* S'ha decidit seleccionar aquest criteri. El raonament per a la selecció de la completesa és semblant a l'explicació donada en el cas de la quantitat, ja que resulta també rellevant que a més de l'obtenció de moltes dades, aquestes siguin completes, ja que si no ho són, aquestes no podràn ser usades perquè no resultaran rellevants en molts dels casos d'estudi. Per a l'assignació d'un valor per aquest criteri, s'usarà l'històric de dades rebudes i es calcularà el tant per cent de N/As existents de mitjana en les dades. Posteriorment, s'invertirà el resultat obtingut i aquesta serà la puntuació final en quant a completesa.

- *Representació concreta.* S'ha decidit descartar aquest criteri. La justificació d'aquesta decisió està plenament basada en la importància d'aquest factor en el resultat final. S'ha vist que el fet de tenir unes dades més o menys compactes no acaba influint en les conclusions obtingudes en els anàlisis sinò que únicament en facilita el tractament; així doncs, com que teòricament es pretén automatizar el tractament, s'ha conclòs que aquest índex no aportaria valor al resultat final del CBR.

- *Representacio coherent.* S'ha decidit seleccionar aquest criteri. Aquest subíndex és important, ja que, de cara a l'automatització del procés de tractament i a l'emmagatzematge de les dades, el fet de tenir un conjunt de dades que sigui coherent en el temps i presenti formats similars resulta un tret determinant. Com més canvis pateixin les dades més s'hauran d'actualitzar els sistemes de tractament i de preprocessament, influint així en el rendiment del sistema a dissenyar. Per a la mesura d'aquesta mètrica es pretén realitzar un estudi dels canvis que han patit les diferents fonts de dades durant el temps a partir de l'històric de dades que es conserva. D'aquesta manera, com més canvis hagi tingut una font de dades, menys puntuació rebrà.

- *Facilitat de manipulació.* S'ha decidit seleccionar aquest criteri. Aquesta selecció es veu justificada per l'increment de valor que tenen les dades en el cas d'estudi si són usables en més d'un context o en diferents tipus d'estudis. Per tant, la facilitat en el tractament de les dades haurà de ser un punt a tenir en compte pel càlcul de l'índex general. La puntuació d'aquest criteri es realitzarà d'acord amb el format de les dades rebudes i en la facilitat estimada assignada per la persona responsable de la realització del sistema de tractament de dades.

- *Llibertat d'errors.* S'ha decidit seleccionar aquest criteri. El fet de seleccionar aquesta mètrica es deu al fet que, com s'ha expressat anteriorment en el punt de credibilitat, existeix una necessitat molt gran d'assegurar que les dades són correctes i s'ajusten a la realitat, per tal de poder oferir una anàlisi de qualitat i no deteriorar la imatge donada per l'empresa. Serà per tant, un criteri molt vinculat al de credibilitat. Per tal d'avaluar la puntuació en aquesta mètrica, es prendrà de nou l'històric de dades i s'analitzarà la quantitat d'errors existents de mitja per cada font de dades. A partir del resultat obtingut, s'assignarà una puntuació representativa per a cada un dels casos.

- *Interpretació.* S'ha decidit descartar aquest criteri, ja que, les definicions de les metadades de les diferents fonts no són accessibles i tampoc usades en el cas d'estudi. Els treballadors fa molt temps que es troben treballant amb les mateixes dades i les entenen basant-se en la seva experiència; per tant, no tindria sentit donar una puntuació que únicament rebaixaria la puntuació final de la qualitat basant-se en fets que ni s'usen en el cas d'estudi ni són representatius de cara al futur dins de l'empresa. Existeix però, la possibilitat que en el futur, es pugui integrar aquest subíndex en cas de resultar necessari si hi ha un canvi de mentalitat entre els treballadors.

- *Objectivitat.* S'ha decidit descartar aquest criteri. La presa de decisió en aquest cas s'ha degut al fet que cal recordar que les dades que s'agafen per aquest estudi són qualitatives, fet que fa que teòricament, parteixin d'un nivell d'objectivitat i siguin comprovables. A més a més, les fonts de dades externes contractades no obtenen cap benefici basat en el contingut de les dades, sinó que l'obtenen a partir de la contractació d'un servei de proporcionament de dades per a l'estudi i que, per tant, són empreses que asseguren aquesta objectivitat buscada. Així doncs, aquest índex no resultava rellevant per al càlcul del CBR.

- *Rellevància i valor afegit.* S'han decidit seleccionar aquests criteris com a agrupació. Aquests dos criteris s'han seleccionat en conjunt ja que, en aquest cas específic, resulta molt important que les dades afegides siguin rellevants i aprotin un valor real a l'estudi. Recordem doncs que en cas d'afegir dades no rellevants, es pot arribar a reduïr la precisió o donar un sobreaprenentatge al model, produint així, un augment de l'error en els càlculs. Per fer el càlcul d'aquests subíndex es mirarà el conjunt de fonts de dades existents i es comprovarà què aporta cada font diferent de les altres. Finalment, amb les observacións obtingudes, es donarà una puntuació basada en la diferenciació entre les fonts de dades.

- *Seguretat.* S'ha decidit descartar aquest criteri. Això és degut a la poca importància que representa aquest factor dins del procés d'anàlisi de les dades. Les dades resulten ser públiques pels clients de les diferents fonts i, per tant, el paper de la seguretat de les mateixes és secundari al no representar dades privades i confidencials. Així doncs, no s'ha seleccionat el criteri ja que, no és un fet que es valori especialment dins del cas d'estudi de cara a la selecció de les diferents fonts de dades i aplicar-lo suposaria expressar un factor no representatiu dins del total del CBR.

- *Puntualitat.* S'ha decidit seleccionar aquest criteri. Aquest és un dels índexs que més problemes donarà entre les diferents fonts de dades perquè pot suposar una pujada molt alta en la puntuació d'alguns i una baixada significativa per altres. Tot i això, aquest factor acaba sent molt rellevant pels estudis números, ja que, la freqüència de la rebuda de dades hauria de ser idíl·licament constant i, si bé és cert que les fonts de dades contractades ho ofereixen, el cas de les dades del client no. Per tant, la utilització d'aquest s'haurà de vigilar, ja que, en cas de desnivellar-lo podria reduir excessivament la qualitat de les dades del client. Per a mesurar-ho, es demanarà als diferents stakeholders interns quina creuen que ha de ser la puntuació assignada en cada cas i es realitzarà la mitjana dels resultats obtinguts.

- *Comprensibilitat.* S'ha decidit no seleccionar aquest criteri, ja que, en molts casos, les dades no resulten comprensibles perquè necessiten un tractament previ. Cal destacar que en estudiar els casos, s'ha vist que en gairebé tots ells, el fet que les dades inicials no siguin comprensibles no representa que després d'un lleuger tractament, aquestes siguin usables i produeixin outputs tant o més vàlids que les dades que des de bon principi eren comprensibles. Així doncs, com que el criteri no afegeix valor al resultat final, s'ha decidit descartar-lo per evitar desviacions no justificables.

Un cop acabada la selecció dels diferents índexs, ens proposem presentar una taula a mode de resum amb els criteris seleccionats i els seus pesos assignats. Per fer l'assignació d'aquests pesos ens hem basat en un acord entre els diferents components del departament de Data Science. Aquest acord consisteix en la utilització d'un mètode semblant a un "Poker Plan"; és a dir, per cada subíndex cada treballador ha plantejat una puntuació entre 0 i 100 i, posteriorment, els treballadors amb la puntuació més alta i la puntuació més baixa respectivament han justificat la seva decisió. Finalment, mitjançant un acord conjunt, s'ha decidit una punutació justa pel criteri. Un cop avaluats tots els subíndexs, s'han realitzat petits canvis per establir un pes total de 100. Així doncs s'ha obtingut la següent taula (Taula 8.1):

Subindex (BCI)	Seleccionat?	Pes (W)
Accessabilitat.	0	-
Quantitat.	1	20
Credibilitat i reputació.	1	10
Completesa.	1	20
Representació concreta.	0	-
Representació coherent.	1	10
Facilitat de manipulació.	0	-
Llibertat d'errors.	1	15
Interpretació.	0	-
Objectivitat.	0	-
Rellevància i valor afegit	1	10
Seguretat.	0	-
Puntualitat.	1	15
Comprensibilitat.	0	-
TOTAL		100

Taula 8.1: Taula de selecció dels criteris i els pesos (W)

Tal com es pot observar a la taula (8.1), s'han acabat seleccionant 9 dels 16 subíndexs proposats. A més a més, veiem que es compleix la condició que afirmava que el pes total havia de ser 100. Finalment, la divisió de pesos ha estat bastant equilibrada entre els diferents índexs; tot i això, s'ha donat una importància alta a la quantitat i completesa de les dades, una rellevància mitjana a la llibertat d'errors i a la puntualitat i un valor lleugerament més baix a la credibilitat, a la reputació, a la representació coherent, a la rellevància i al valor afegit. Tot i això tan sols existeix una diferència d'un 5 per cent entre els nivells de rellevància i un màxim d'un 10 per cent de distància entre nivells donada l'existència d'únicament 3 agrupacions de puntuacions.

8.3 Anàlisi de les fonts de dades existents

Un cop donada una descripció general de les diferents fonts de dades, a partir de la seva classificació en globals i específiques, i definit el sistema d'avaluació de la qualitat de les dades, procedim a realitzar una anàlisi exhaustiva de cada una de les fonts de dades en ús actualment. Per tal de fer-ho, es dividirà l'explicació en els 4 punts següents:

- *Descripció introductoria:* En aquesta primera secció es donarà una visió genèrica de la font de dades utilitzada per poder situar-nos davant del tipus de font de dades. Per tal de fer-ho, s'usarà tant la informació donada per l'empresa gestora de la font com informacions alienes a aquesta.
- *Metadades:* Una vegada situats, es presentarà una taula explicativa de tots els camps que les fonts de dades a tractar ens proporcionen. Aquesta taula de descripció de les dades es defineix com a fitxer de Metadades (conceptes apresos a Mineria de Dades [MD]) i ens servirà per entendre els camps disponibles i per determinar si dos camps amb el mateix nom o similar en diferents fonts poden ser creuables o es refereixen a conceptes totalment diferents. A més a més, ens facilitarà la comprensió en profunditat de les dades. Resulta de vital importància doncs, realitzar un fitxer de metadades el màxim d'exhaustiu possible per tal de, posteriorment, poder justificar que els encreuaments realitzats són coherents.
- *Avaluació qualitat:* Posteriorment a la descripció de la font de dades, es realitzarà una anàlisi de la qualitat de les dades basada en l'aplicació del sistema d'índexs CBR explicats anteriorment. Aquesta part ens permetrà avaluar i justificar la qualitat de les dades a utilitzar, assegurant així que les dades extretes de les diferents fonts aporten valor i són correctes per al seu respectiu ús en les anàlisis.
- *Diagrama UML:* Finalment, es presentarà un diagrama UML (estudiat a IES, BD, CBD...) de representació en forma de classes de les diferents fonts de dades. Aquest punt es realitza per tal de facilitar la comprensió del futur funcionament del sistema d'emmagatzematge d'aquestes i per facilitar la visualització de possibles similituds en les estructures de les diferents fonts de dades extretes. Una idea de futur en cas de ser possible la seva realització, consisteix en la integració de tots els diagrames UML generats per tal d'especificar una estructura de dades que doni cabuda a totes les fonts.

Comencem doncs a analitzar les diferents fonts de dades, iniciant per les fonts de dades globals, de les quals es podrà realitzar una anàlisi més concreta degut a la seva estabilitat i ús conjunt.

8.3.1 Global - Instar Analytics i Kantar Media [KM]

Iniciem la descripció introductòria d'aquesta font de dades definint què és Instar Analytics i Kantar Media. No trobem millor manera de fer-ho que començar oferint la definció que ens proporcionen ells mateixos dels seus propis serveis:

> **Citació 3** *Ïnstar Analytics is a flexible tool.*
>
> - *An app accessible on smartphones and tablets which enables on the move access to get a snapshot of how programmes and ad spots have performed*
> - *Web and desktop versions for deep-dives into the data and granular analysis*
> - *An API enabling data to be delivered into third party applications.*
>
> *The software integrates completely between audiences, programmes and spots. It can run multi-market and multi-media analysis. It can be scaled up to cover a broad range of data, or it can focus on specific data sets.""[20]*

> **Citació 4** *"Kantar Media is a global leader in media intelligence, providing clients with the data they need to make informed decisions on all aspects of media measurement, monitoring and selection. Part of Kantar, the data investment management arm of WPP, Kantar Media provides the most comprehensive and accurate intelligence on media consumption, performance and value."[21]*

Ens trobem doncs davant de dos serveis diferents. Veiem clarament en la definció oferida que Instar Analytics és un software que ens permetrà visualitzar les dades i realitzar l'extracció de les mateixes mitjançant una eina visual o mitjançant una API mentre que, Kantar Media tracta únicament el conjunt de dades que l'empresa ofereix i que s'usen per a l'estudi analític posterior. Com a resum doncs veiem que la font de dades real és Kantar Media, i que Instar Analytics acaba essent una eina que ens permet l'explotació d'aquesta. Així doncs, a partir d'aquest moment, es tractarà la font amb el nom de Kantar Media [KM], ja que, resulta ser la font de dades explícita.

Un fet rellevant en la definició donada és que Kantar Media fa referència a WPP, conglomerat d'empreses publicitàries del que s'ha parlat en la definició de context inicial, recordem que aquest és el conglomerat més gran en el sector publicitari, fet que ens pot ajudar de cara a justificar alguns dels criteris establerts per al càlcul del CBR.

Un cop diferenciats els dos conceptes es pretén definir les dades que realment ofereix la font de Kantar Media. Amb l'objectiu d'assolir aquesta definició, tornem a citar l'auto explicació que ofereix el mateix servei.

> **Citació 5** *"Our data is key to identifying, understanding, reaching, refining and measuring audiences as well as understanding context and trends across media. Our datasets include: audience profiles for targeting stated, purchase data, inventory availabilities and rates, paid search keywords, paid search spend, paid ad occurrences, paid ad spend, earned media occurrences. brand/issue sentiment scoring, estimated media value (EMV), audience and media ratings. All of our data uses currency grade methodology and crosses on and offline media to provide a more holistic view."[22]*

Deixant de banda les especificitats, veiem que en general aquesta font de dades ens aporta la informació

de la mesura de les audiències donant també el context d'aquestes. A més a més, ens parla que aquesta informació no és tan sols referent a un mitjà, sinó que fa referència a un encreuament de tots els mitjans, fet que la fa molt més valuosa en representar més genèricament l'impacte publicitari. Un altre punt rellevant, i passant ara a les dades específiques que ens ofereix, és que ens dóna una desagregació de les audiències per perfils, informació de la compra, informació dels diferents spots i les diferents mètriques de mesura de l'impacte publicitari. Concloem doncs que estem tractant amb una font de dades basada en la mesura d'audiències segons les definicions generades per l'empresa gestora de la font.

Procedim ara a parlar amb l'equip per veure l'ús real que se n'està fent de la mateixa i fins a quin punt aquestes dades resulten útils. De la reunió realitzada n'extraiem la següent informació:

1. L'equip utilitza les dades en el context televisiu per a calcular l'impacte que tenen els diferents spots contractats en els diferents targets i àmbits rellevants per a l'estudi. Per tal de fer-ho, s'utilitza una sola mètrica anomenada GRPs i totes les dades de context amb l'objectiu de tenir el màxim de flexibilitat de cara a les possibles agrupacions de dades necessàries.

2. Existeix la possibilitat d'afegir el nombre de mètriques tractades en un futur.

3. Kantar Media resulta una font molt fiable, ja que els preus de les diferents accions publicitàries es basen en els resultats oferts per aquesta plataforma. Més concretament, el preu d'un acte publicitari es basa en els GRPs que apareixen per aquest a la font de Kantar Media. Aquest fet, és indicatiu que és una font amb molta reputació i fiabilitat, ja que al final les dades d'aquesta font de dades tindran influència directa en el preu pagat per cada acte publicitari.

4. Kantar Media és una font molt actualitzada. Rep actualitzacions i correccions de dades mal introduïdes cada dia, fet que fa que, les dades siguin d'una gran fiabilitat i no continguin gairebé cap error perquè aquests són corregits en aproximadament un o dos dies vista.

5. Les dades extretes d'aquesta font, són la base de les anàlisis realitzades, ja que, aporten granularitat diària (a nivell spot), cosa que permet realitzar anàlisis molt exhaustives i donar ajustos a la realitat molt més elevats que les altres fonts a tractar.

Concloem doncs veient que l'ús que s'està donant actualment a IKI Media Communications S.L. de la font de dades Kantar Media, s'ajusta amb el servei ofert per aquestes, ja que, principalment s'usa per a l'estudi d'audiències. A més a més, d'aquesta primera definició n'extreiem moltes dades útils de cara a l'estudi de la qualitat com podria ser la reputació de la font, la baixa taxa d'errors que es troben en aquest servei o la freqüència d'actualització de les dades.

Una vegada situats en l'utilitat d'aquesta font, presentem la taula de Meta Dades que ens permetrà acabar d'entendre les dades reals que ens ofereix aquest servei. Per tal de facilitar-ne el tractament, a la taula de meta dades tan sols s'han introduït les dades que s'usen en els estudis analítics; és a dir, les dades referents a la definició del context i les referents a la mètrica dels GRPs amb els diferents targets i àmbits. (Taula 8.2 i Taula 8.3)

Atribut	Tipus	Descripció	Exemple
Sector	String	Sector en el qual es troba el producte anunciat	BELLEZA e HIGIENE
Categoria	String	Categoria que te dins del sector assignat	Productos CABELLO
Producto	String	Línia de productes dins de la categoria assignada	Fijadores y Moldeadores
Anunciante	String	Anunciant que paga per l'anunci del producte	PROCTER and GAMBLE ESPAÑA, S.A.
Marca	String	Marca a la que representa l'anunciant	PANTENE PRO-V
Modelo	String	Model que pretén promocionar la marca amb l'anunci	ESPUMA RIZOS
Campaña	String	Campanya publicitaria a la qual pertany l'anunci, sovint combinació de marca + model	PANTENE PRO-V/ESPUMA RIZOS
Cadena	String	Cadena en la que es visualitza l'anunci	TV3
Fecha	Date	Dia en el que s'emet l'anunci	20/02/1996
Hora de Inicio	Hour	Especifíca l'hora a la qual comença l'anunci	12:40:15
Duracion	Integer	Especifíca la durada de l'anunci en segons	0000:20
Precio	Integer	Preu de l'espai publicitari en euros	2600
Tipo	String	Tipus d'anunci que s'emet	NORMAL
Formato	String	Format de l'anunci que s'emet	SPOT NORMAL
Contingut	String	Estil de l'anunci que s'emet	PUBLICIDAD CONVENCIONAL
Context	String	Defineix si un anunci ha estat donat en pantalla completa o solapat	NO SOLAPADO

Taula 8.2: Kantar Media metadata (Part 1)

Atribut	Tipus	Descripció	Exemple
Tipologia	String	Tipologia de l'anunci que s'emet	SPOT
Pantalla compartida	Boolean	Defineix si s'emet l'anunci amb la pantalla compartida amb altres imatges	NO SOLAPADO
Promocion	String	Promoció realitzada durant el període d'emissió de l'anunci	Spot
Pos. Bloque 1	Integer	Posició en la qual es troba sobre el bloc d'anuncis entre un programa i un altre (càlcul 1)	3
Spots Bloque 1	Integer	Nombre d'anuncis entre un programa i un altre (càlcul 1)	4
Pos. Bloque 2	Integer	Posició en la qual es troba sobre el bloc d'anuncis entre un programa i un altre (càlcul 2)	3
Spots Bloque 2	Integer	Nombre d'anuncis entre un programa i un altre (càlcul 2)	4
Pos. Bloque 3	Integer	Posició en la qual es troba sobre el bloc d'anuncis entre un programa i un altre (càlcul 3)	4
Spots Bloque 3	Integer	Nombre d'anuncis entre un programa i un altre (càlcul 3)	5
Titulo Emision	String	Nom de l'emissió en la qual apareix l'anunci	EL PROGRAMA DE ANA ROSA
Descripcion creatividad	String	Identificador de la creativitat assignada a l'anunci	2016/07/01 0042
GRPS	Double	Especifíca el nombre de GRPs per anunci (1 x Target x Dimensió)	5.45

Taula 8.3: Kantar Media Metadata (Part 2)

Queda doncs definida la taula de meta dades (Taula 8.2 i Taula 8.3), on es poden veure els diferents camps que ofereix la font de dades Kantar Media. És observable que gairebé tots els camps que ofereix serveixen per contextualitzar l'anunci i que l'única mètrica que s'usa és el GRP, expressat en forma de diferents columnes, una per a cada target (t) i àmbit existent (a) (t x a). Aquest fet ja es deixava entre veure en les explicacions donades pels treballadors, que parlaven d'atributs d'agrupació i mètrica de GRPS, què és exactament el resultat que extraiem de les metadades. Observem també que existeixen tres camps referents a la posició de l'anunci i el nombre de spots en l'espai publicitari. Certament, aquests tres camps expressen un mateix contingut però usant diferents mètodes de càlcul, diferenciant-se principalment en la comptabilització dels anuncis vinculats a la cadena.

Acabant ara amb la definició de la taula de meta dades i passant ara, tal com prèviament s'ha plantejat, a l'avaluació de la qualitat, s'ha calculat la següent taula amb les puntuacions i justificacions de les mateixes per tal d'obtenir el valor final de l'índex CBR. (Taula 8.4)

Subindex(BCI)	Justificació	Pes (W)	Avaluació	Aportació
Quantitat	Les dades són donades a nivell de segon, per tant, la quantitat de les dades és la més alta possible i ha de rebre el màxim de puntuació.	20	**10**	2
Credibilitat i reputació	Degut a que aquesta eina és usada tant per WPP i les mètriques són acceptades per les empreses de mitjans la reputació i credibilitat es la més elevada. Recordem que les empreses de mitjans basen la facturació en aquestes dades.	10	**10**	1
Completesa	En aquestes dades o no existeixen N/As o en els 5 arxius analitzats no se'n ha trobat cap, per tant la punutació final torna a ser la més elevada.	20	**10**	2
Representació coherent	Els canvis en aquest datasource són gairebé nuls. A més a més, quan és donen, és manté durant un temps provisional la versió antiga per a facilitar el canvi del sistema.	10	**9**	0,9
Llibertat d'errors	Els errors son existents però al existir correccions diàries, és redueixen al llarg del temps. Aixi doncs no és pot calcular l'index de la manera indicada perquè el nombre d'errors no és estàtic. Per tant s'assigna de manera subjectiva i sabent que els errors són pocs i es corregeixen	15	**8,5**	1,275
Rellevància i valor afegit	Com ja s'ha dit, és la font més rellevant i què mes valor aporta als anàlisis. Aporta les mètriques.	10	**10**	1
Puntualitat	Les dades són actualitzades diàriament, per tant, són puntuals i molt freqüents.	15	**10**	1,5
TOTAL			**9,68**	

Taula 8.4: CBR per Kantar Media [km]

Veiem a la taula 8.4 que la puntuació obtinguda en l'avaluació de la qualitat és de 9.68. Fet que no és sorprenent, ja que sabem que aquesta és la font de dades que més valor, reputació i freqüència aporta a les anàlisis. De fet, és la font de dades més usada en el món de les agències de mitjans i, per tant, resulta lògic que obtingui una puntuació molt elevada. A més, és una font de dades acceptada per la comunitat i amb una llarga experiència en el sector. Així doncs, aquesta font de dades té la qualitat necessària per a ser usada en els estudis.

Una vegada donada una puntuació de la qualitat de la font de dades KM, passem a generar l'últim apartat d'aquesta anàlisi de la font de dades, és a dir, procedim a dissenyar l'UML de l'estructura de classes.

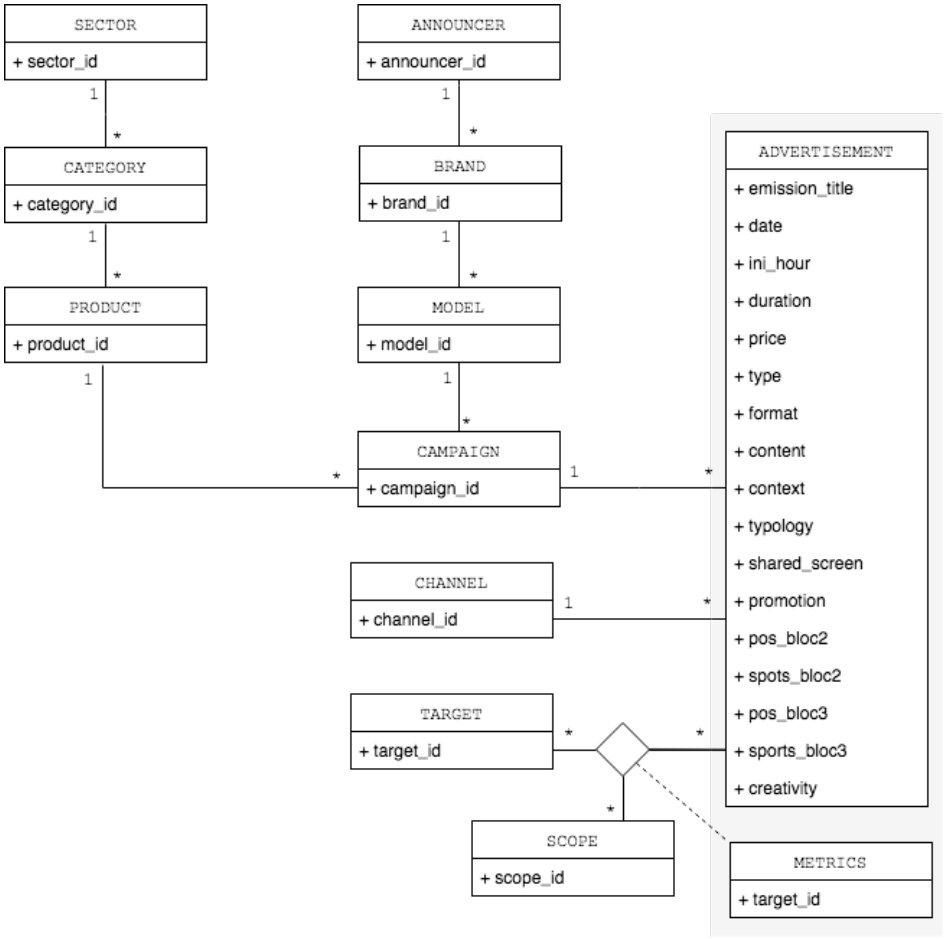


Figura 8.3: Diagrama de classes UML per a Kantar Media [km]

S'ha donat aquest disseny UML, per tal de definir l'estructura de les dades que es reben. La decisió de separar sector, categoria, producte, anunciant, marca, model i producte en diferents classes és deguda al fet que, en l'àmbit intern, són els atributs que s'utilitzen per encreuar i agrupar principalment. A més a més, existeix una relació entre aquests camps, que es veu expressada en la taula de meta dades. Així doncs, tenim una estructura de jerarquies representant aquests atributs convergents a l'atribut campanya, classe que manté un vincle amb l'objecte anunci que conté les dades específiques d'aquest. Finalment, existeix una relació ternària de target, àmbits i anunci que ens permet generar una classe mètrica que serà la que mantingui les mètriques a usar. S'ha decidit crear una classe mètrica en comptes d'una classe GRP de manera que, en cas de voler incrementar el nombre de mètriques (fet no descartable com prèviament ha estat expressat) es pugui afegir la informació a l'objecte Mètrica sense que el nom deixi de representar el contingut.

Donem per conclòs l'anàlisi de la font de dades KM després d'haver realitzat tots els apartats plantejats per a aquesta font. Deixem doncs aquí tota la informació referent a aquest punt per al seu futur ús durant la fase d'anàlisi de requisits i disseny de la solució.

8.3.2 Global - Info Adex [INV]

Donem pas a l'anàlisi de la font de dades Info Adex mitjançant una definició introductòria que seguirà una estructura molt semblant a la realitzada anteriorment per KM. Comencem doncs presentat el punt de vista que aporta la mateixa empresa encarregada de la distribució de les dades sobre el tipus de font de dades que són i trobem que:

> **Citació 6** *"En InfoAdex disponemos de una base de datos online en la que nuestros clientes podrán encontrar toda la información cuantitativa relacionada con la actividad publicitaria de España. Debido a la manutención diaria de la base de datos y a la información que ésta es capaz de proporcionar, se pueden realizar análisis de distinta índole, abarcando algunos temas como: análisis de la competencia, planificación de campañas, definición de estrategias de marketing o la preparación de briefings. Esta información se captura a través de todos los anuncios que aparecen en televisión, radio, diarios y sus suplementos, dominicales, revistas, publicidad exterior, Internet y cine. Nuestro trabajo es registrar todos y cada uno de estos anuncios, validarlos y ponerlos a disposición de nuestros clientes."[23]*

Com podem veure, la font de dades Info Adex és una font d'informació quantitativa relacionada amb l'activitat publicitària Espanyola. A diferència de la font de Kantar Media però, aquesta ens aporta informació recopilada de tots els mitjans, és a dir, de televisió, ràdio, diaris, revistes, publicitat exterior, internet i cinema. Aquest fet fa que existeixi una diferència notòria amb el proporcionat per la font de Kantar Media que tan sols està aportant informació de televisió (tot i estar tenint en compte tant la televisió online com la offline). A més, segons la seva pròpia informació, aquestes dades que proporcionen són prèviament validades i, per tant, certes i sense errors. Aquest fet, ens fa plantejar-nos si aquesta font aporta les mateixes dades que Kantar Media, ja que, en cas de ser així, hauríem de plantejar-nos fins a quin punt resulta important la font de dades de Kantar Media si, Info Adex per si sola ens pot aportar les mateixes dades però més completes. Així doncs, passem a definir les dades que proporcionen, per veure fins a quin punt s'encavalquen amb les dades prèviament exposades. Per fer-ho, tornem a citar el que el mateix proveïdor facilita:

> **Citació 7** *"Proporciona los datos que necesitas para contestar estas y otras muchas preguntas sobre la inversión publicitaria, ofreciéndote también datos de inserciones y ocupación."*

Denotem doncs, després d'aquesta cita, que existeix una diferència clara entre les dues fonts, ja que, Info Adex tan sols ens mostrarà informació relativa a les inversions publicitàries generades en els diferents mitjans i, per tant, no podrem obtenir dades sobre les audiències a partir d'aquesta base de dades. Així doncs, aquesta font ens permetrà rebre informació completa sobre la inversió dels diferents clients en tots els mitjans (cal remarcar el fet que sigui en tots els mitjans perquè en cas que tan sols fos els de televisió, no seria necessari l'ús d'aquesta font, ja que, KM ens ofereix el preu per anunci i, per tant, a partir d'una agregació d'anuncis podríem arribar a saber la inversió televisiva total de cada sector). Justifiquem doncs la necessitat d'aquesta font de dades dedicada específicament a la inversió i anem una mica més enllà per veure què més se'n pot extreure per aprofitar-la al màxim. En aquest punt, descobrim que a part de tot això, aquesta font ens permetrà vincular les inversions realitzades amb les creativitats assignades a cada una de les accions publicitàries existents tal com podem veure en la següent citació:

Citació 8 *"Permite el acceso a la vinculación de cada inserción con su creatividad en diarios, dominicales, internet, radio, redes sociales, revistas y televisión, siempre que tengas contratado el módulo de creatividades."*

Per tant, a més d'oferir la gestió de les dades d'inversió general, en aquesta font podrem consultar les creativitats reals assignades a cada acte publicitari i no tan sols l'identificador assignat a cada una d'elles (a diferència del cas de Kantar Media). Aquest fet, també resulta rellevant perquè, tot i no ser un element a posar en les anàlisis, sovint resulta important la visualització de tots els materials usats durant una campanya publicitària, ja que, depenent de la qualitat i el tipus de material, es pot arribar a justificar de manera qualitativa el perquè de l'efecte publicitari sobre alguns dels targets. Per exemple, trobem el cas de l'anunci de Mixta del Çerdo Volador"; aquest anunci, al tenir un material entretingut i que s'enganxava, amb una inversió menor a el normal, va arribar a més gent pel boca a boca, fet que no era visible en les dades de Kantar Media, ja que no depenia del nombre de visualitzacions purament televisives sinó de l'impacte social que va generar.

Passem ara, com s'ha realitzat en el cas anterior, a intentar trobar la utilitat real per la qual s'està utilitzant a IKI Media Communications S.L. mitjançant una reunió amb els stakeholders interns de l'equip de Data Science. D'aquesta reunió se n'extreu que:

1. L'equip usa les dades extretes d'Info Adex per tal de tractar les inversions generals en les anàlisis.
2. L'equip usa les dades que s'extreuen d'Info Adex per a realitzar una comprobació de les inversions de Kantar Media.
3. Tota l'empresa usa el sistema d'extracció de dades per portar un control de les inversions generals realitzades en les campanyes gestionades per IKI Media Communications S.L.
4. Tota l'empresa usa el sistema per a visualitzar el material usat per a les accions publicitàries. Especialment l'equip de planificació i de data science.

Comprovem una vegada més que els usuaris d'IKI Media Communications S.L usen un subconjunt de la utilitat proposada pel proveïdor de les dades. Aquest fet ens porta a pensar que, probablement, i per un problema ja identificat en el primer estudi de context (portat a terme com a introducció del treball), si es redueix la càrrega de feina dels treballadors, l'explotació d'aquestes dades incrementarà i s'usaran de manera més completa i eficient. Això també ens porta a poder donar per vàlida la definició del proveïdor de les dades.

Val la pena remarcar l'ús real que s'està donant a la funcionalitat de visualització de les diferents creativitats, que, tot i que en un principi pot arribar a semblar irrellevant, acaba oferint un servei extra que redueix de manera significativa el temps. A més, afegeix un valor qualitatiu tant a les anàlisis com als serveis oferts per l'empresa que, molt probablement, en cas de no ser accessibles amb la facilitat que aporta Info Adex, no podria donar.

Queda doncs definida la utilitat i les funcionalitats oferides per aquesta font de dades i, per tant, com a conseqüència, també es dóna per acabada la introducció d'aquesta font i es passa a la realització de la taula de Metadades d'aquesta. Aquesta es presenta a continuació (Taula 8.5):

Atribut	Tipus	Descripció	Exemple
Sector	String	Especifíca el sector en el que és troba el producte anunciat	BELLEZA E HIGIENE
Categoria	String	Especifíca la categoria dins del sector del producte anunciat	PRODUCTO CABELLO
Producto	String	Especifíca la linea de productes dins de la categoria del producte anunciat	ACONDICIONADORES Y SUAVIZANTES
Anunciante	String	Especifíca l'anunciant que paga per l'anunci del producte	PROCTER and GAMBLE ESPAÑA, S.A.
Marca directa	String	Especifíca la marca principal a la que representa l'anunciant	PANTENE
Marca	String	Especifíca la sub-marca a la que representa l'anunciant (Gairebé sempre resulta ser Marca == Marca Directa)	PANTENE
Modelo	String	Especifíca el model que preten promocionar la marca amb l'anunci	PRO V REPA.PRO.
Medio	String	Especifíca el mitja pel qual és transmet	INTERNET
Soporte (x3)	String	Especifíca el suport a partir del qual ha estat emés	20minutos.es
Año	Int	Identifica el any del qual les dades s'han recollit	2016
Mes	String	Identifica el mes del qual les dades s'han recollit	ABRIL
Inversion pondera-da	Double	Valor total de la inversión ponderada mensual	76,193
Inversion	Double	Valor total de la inversión mensual	80,123
Inserción	Double	Valor total de la inserción mensual	0,923

Taula 8.5: Info Adex metadata

A partir de la realització d'aquesta definició dels paràmetres existents en la font de dades, descobrim que aquesta resulta ser molt més senzilla que la font anteriorment tractada (KM). A més a més, ens adonem també que, existeixen camps que tenen el mateix nom i definició entre les fonts, fet que ens permet gairebé assegurar que a partir d'aquests camps, molt probablement, podrem realitzar l'encreuament de les dades de les dues fonts treballades. Fruit d'aquest estudi també trobem nous camps d'agrupació com poden ser mitjà o suport. Un altre punt a tenir en compte és la granularitat de les dades, en aquest cas, aquesta resulta ser mensual i per tant, existirà una quantitat menor de dades d'inversió total.

Una vegada definides les meta dades de la font de dades d'Info Adex, passem a definir amb la mateixa taula usada anteriorment l'índex CBR de la font per determinar-ne la qualitat. Queden definides a la Taula 8.6.

Subindex(BCI)	Justificació	Pes (W)	Avaluació	Aportació
Quantitat	Les dades són actualitzades mensualment, per tant, és troben en una freqüencia baixa i han de tenir una mala punutació.	20	**5**	1
Credibilitat i reputació	Aquesta eina mitjançant una cerca ràpida sembla tenir una reputació bona. A més a més és un datasource prou usat en el sector per la capacitat de centralitzar totes les inversions.	10	**8**	0,8
Completesa	En aquestes dades o no existeixen N/As. El fet de que moltes empreses gestioni l'inversió publicitaria a partir d'aquestes dades fa que sempre es trobin completes i en cas de no ser-ho, hi hagin queixes i siguin automaticament actualitzades.	20	**10**	2
Representació coherent	Els canvis en aquest datasource són gairebé nuls. Però quan existeixen, és realitzen de cop i poden suposar un problema fins a la completa adaptació.	10	**7**	0,7
Llibertat d'errors	Els errors son existents però al existir correccions diàries, és redueixen al llarg del temps. Aixi doncs no és pot calcular l'index de la manera indicada perquè el nombre d'errors no és estàtic sinò decremental. Per tant s'assigna de manera subjectiva.	15	**8,5**	1,275
Rellevància i valor afegit	Com ja s'ha dit, és una font rellevant aprotant valors d'inversió generics i visualització de creativitats.	10	**8**	0,8
Puntualitat	Les dades són actualitzades mensualment, per tant, no són gaire freqüents i sovint la arriben 2 dies fora de plaç	15	**7**	1,05
TOTAL			**7,63**	

Taula 8.6: CBR per InfoAdex [inv]

Observem a la taula 8.6 que la puntuació final és de 7.63. Aparentment, ens trobem davant d'un valor alt de qualitat de les dades tot i que més baix que l'anteriorment tractat (de 9.68). Aquest fet és degut, tant a què aquesta font de dades té una credibilitat inferior a l'anterior com a què el valor de les dades aportades és inferior, sigui per la menor quantitat o per la menor influència d'aquestes. Tot i així, l'avaluació obtinguda en l'índex CBR no ha estat baixa i per tant, podem afirmar que la qualitat de les dades d'aquesta font està assegurada.

Una vegada donada una puntuació de la qualitat, passem a generar l'últim apartat d'aquesta anàlisi de la font de dades, és a dir, procedim a dissenyar l'UML de l'estructura de classes.

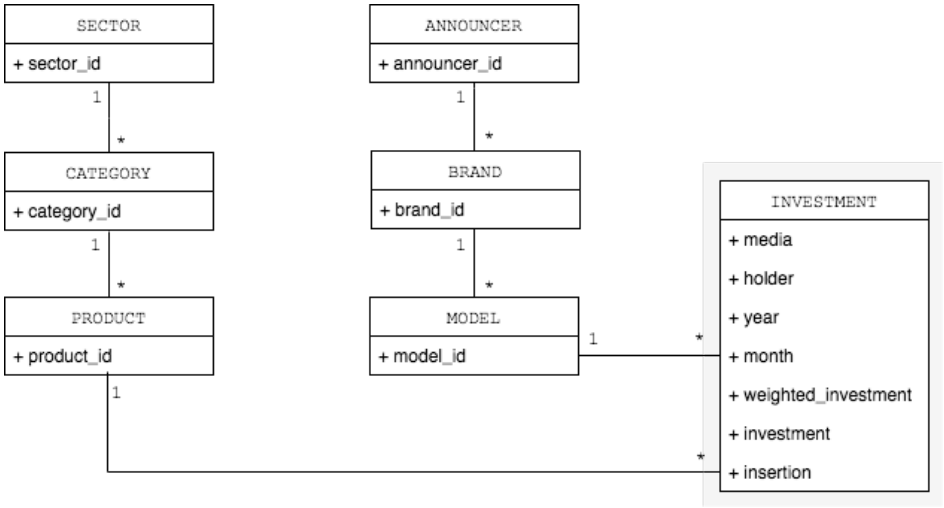


Figura 8.4: Diagrama de classes UML per a InfoAdex [inv] (v1)

Podem veure a la (Figura 8.3) l'UML que s'ha donat com a solució d'estructura de classes per a la font Info Adex. En aquesta estructura, veiem que existeixen uns objectes que segueixen el mateix format que els presentats per a Kantar Media, essent aquests: sector, categoria, producte, anunciant, marca i model. Tot i estar representant el mateix (conceptualment) cal remarcar que els camps pertinents a aquests objectes no tenen per què ser exactament iguals entre les dues bases de dades; per exemple, podem trobar-nos amb què una guardi el valor "NIKEï l'altre el valor "Nike"que aparentment representen el mateix però no són iguals. Aquesta estructura jeràrquica generada, s'acaba vinculant pels seus dos extrems amb una altra classe d'objecte que manté la informació mensual de la inversió ponderada, la inversió i la inserció agrupada segons el suport i el mitjà. Un esquema UML alternatiu i generable a partir d'aquest seria un que suposés Suport i Mitjà com a objectes i no com a atributs d'una classe com el següent (Figura 8.4)

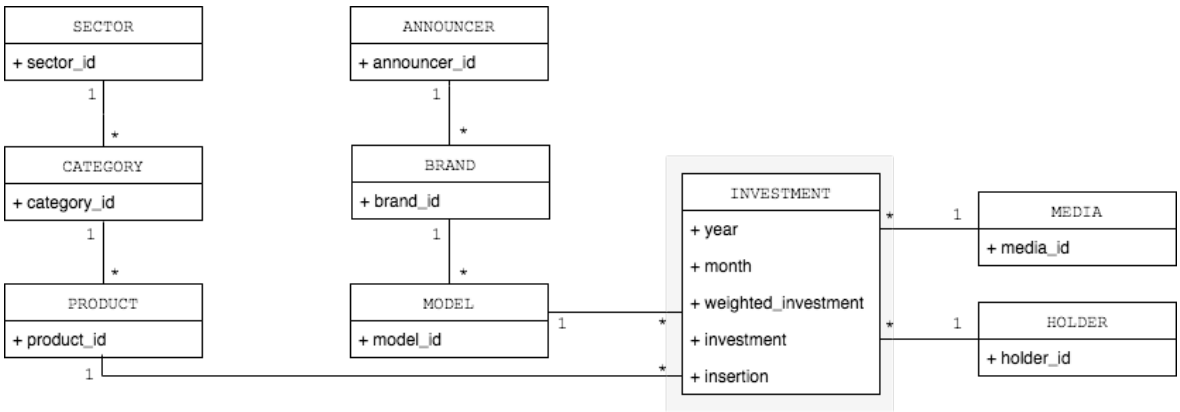


Figura 8.5: Diagrama de classes UML per a InfoAdex [inv] (v2)

Donem per conclosa l'anàlisi de la font de dades INV després d'haver realitzat tots els apartats plantejats per a aquesta font. Deixem doncs aquí tota la informació referent a aquest punt per al seu futur ús durant la fase d'anàlisi de requisits i disseny de la solució.

8.3.3 Global - Kantar TNS [IOPE]

Iniciem l'anàlisi de l'última font de dades anomenada TNS Kantar. Altre cop, per tal de fer la descripció introductòria, tractarem primerament amb la informació oferida pel mateix proveïdor del servei, essent aquest altra vegada part de Kantar en una de les seves filials TNS. La definició oferida és la següent:

> **Citació 9** *"El Tracking IOPE de TNS mide la notoriedad publicitaria de todas las marcas y productos en 8 medios publicitarios distintos, -TV, periódicos, revistas, radio, cine, internet, exterior y publicidad directa- y la contribución de cada medio a la creación de notoriedad de cada marca o producto. Todas las marcas parten en condiciones de igualdad ya que las preguntas son totalmente abiertas y las respuestas son absolutamente espontáneas, porque no se sugieren productos ni marcas.El Tracking IOPE permite...*
>
> - *Medir la eficacia publicitaria: compara semanalmente la notoriedad alcanzada con el esfuerzo publicitario realizado (GRP's, inversiones,...).*
> - *Evaluar las campañas propias y de la competencia*
> - *Tener un post-test permanente de la propia marca y de la competencia*
> - *Definir objetivos de notoriedad y asignar los recursos adecuados a partir de la experiencia adquirida en anteriores trackings."*

De la definició del proveïdor observem que el Traking IOPE de Kantar TNS manté informació sobre la notorietat publicitària de 8 mitjans publicitaris diferents. Fet que, encaixa directament amb les dades ofertes per Info Adex. Així doncs, s'ha de seguir analitzant per trobar la diferència entre les dues fonts usades o per poder justificar la igualtat de les dues per tal de posteriorment plantejar la retirada d'una d'elles. Seguim llegint la definició i trobem la primera diferenciació entre les dues fonts en la freqüència d'actualització i pujada de dades que, en aquest cas, és setmanal mentre en el cas d'Info Adex resulta ser mensual. També ens trobem que, en aquesta font, no es realitza un tractament explícit de la creativitat de l'acció publicitària i, per tant, aquest fet pot ser determinant per a justificar la diferenciació. A més a més, també veiem que en aquest cas les dades són més agregades que en el cas anterior, ja que, com posa a l'inici de l'explicació donada, les dades són per marca i producte, mentre que les altres dades estan classificades per sector, categoria, producte, anunciant, marca i model. A part dels punts prèviament explicats, no trobem més diferències entre les dues fonts. Per tant, valdrà la pena en aquest cas demanar a l'equip una justificació real de l'ús de les dues fonts de dades per veure'n la necessitat justificada o simplement descartar-ne una de les dues.

Realitzem una reunió com les fetes anteriorment per tal de demanar a l'equip l'ús real que es fa de la font de dades i demanem una explicació clara de la diferència d'ús entre aquesta font i la font oferida per Info Adex. D'aquesta reunió n'extraiem la següent informació:

1. La font de dades s'usa per a realitzar un seguiment de la inversió realitzada a curt termini (setmanalment).
2. La font de dades s'usa per a estudis analítics amb més granularitat temporal.
3. La font de dades, al ser més agregada i oferir una confiança inferior a l'oferta per Info Adex, no s'usa per a la facturació final.
4. Resulta massa agrupada i per tant, es desitjaria trobar un sistema per donar una major desagregació de les dades.

5. Permet el filtre de les dades segons els targets existents, permetent així poder comprovar les inversions dedicades a cada mercat objectiu.
6. Aquest sistema no permet la visualització del material usat per a l'acció publicitària.

Després de la realització d'aquesta reunió extraiem tant la visió de l'equip d'aquesta font com la justificació de diferenciació necessària. Primerament, veiem que aquesta font s'usa per a gestionar la inversió a escala setmanal de les diferents marques amb les quals es treballa. També, se n'extreu la importància que es dóna que aquesta inversió pugui ser desagregada en funció dels targets objectiu, fet que fa que resulti molt interessant de cara a estudis estadístics, ja que, poder discretitzar entre els mercats objectius acaba permetent oferir al client una visió de la inversió dividida entre els diferents targets impactats, podent justificar així el compliment dels dividends proposats en la planificació.

Per acabar amb la descripció inicial, s'ha demanat la justificació de la diferenciació i s'ha obtingut la següent resposta: L·La principal diferenciació, és que mentre que una és més fiable i ofereix una major desagregació, l'altra ofereix més freqüència i desagregació per targets. Aquest fet fa que ambdues s'acabin usant amb objectius diferents depenent del cas d'estudi. A més a més, Info Adex ofereix la possibilitat de visualitzar el material publicitari mentre que en el cas de Kantar TNS no s'ofereix". Per tant, podem concloure que ambdues fonts de dades actualment resulten importants i rellevants per als estudis, tot i que s'haurà de mantenir un control perquè en cas que alguna de les dues millorés podrien arribar a ser redundants (ja que actualment les diferències entre les dues no són immenses).

Concloem doncs la descripció introductòria de la base de dades i passem a definir la taula de meta dades, fet que ens permetrà entendre millor els camps oferts i acabar de tancar les diferències entre les fonts. La taula es presenta a continuació:

Atribut	Tipus	Descripció	Exemple
Sector	String	Especifíca el sector en el que és troba el producte anunciat	MEDIOS DE COMUNICACION Y TELECOMUNICACIONES
Marca	String	Especifíca la marca principal a la que representa l'anunciant	RASTREATOR
Medio	String	Especifíca el mitja pel qual és transmet	Televisión
Semana	Int + String	Setmana a sobre la que tracten les dades. Expresada tant en nombre com en rang de dates.	1 [31/12/2012-06/01/2013]
Notorierad Semanal	Double	Dada de la notorietat setmanal (1 x Target)	5.3

Taula 8.7: Kantar TNS metadata

Un cop introduïda la taula 8.7, podem veure que, resulta clarament un conjunt de dades amb molts menys atributs que les ofertes anteriorment tant per Info Adex com per Kantar Media. Aquest fet no resulta sorpenent, ja que, sabíem que les dades resultaven ser molt més agregades que les ofertes per les altres fonts de dades. Queda clar doncs que aquesta font de dades l'únic que acaba oferint per sobre de Info Adex és la discretització per targets i una freqüència superior però que, a nivell general, acaba aportant la mateixa informació. Resulta important comentar que el fet que aquesta base de dades sigui del mateix proveïdor que la de KM, comporta que els camp d'agrupació, sector i marca siguin exactament iguals en les dues fonts i que, per tant, permetin un creuament directe sense necessitats de conversions intermèdies tot i que aquest fet no hauria de resultar rellevant, ja que, Kantar Media ja ofereix un sistema de càlcul d'inversió televisiva com s'ha comentat anteriorment. Pel que fa a la vinculació amb INV, cal remarcar que a diferència del que passa amb KM, es pot assegurar que els camps expressen un mateix concepte però no que ho expressin amb un mateix format.

Una vegada definides les meta dades de la font de dades de Kantar TNS, passem a definir amb la mateixa taula usada anteriorment l'índex CBR de la font per determinar-ne la qualitat. Queden definides a la taula 8.8.

Subindex(BCI)	Justificació	Pes (W)	Avaluació	Aportació
Quantitat	Les dades són actualitzades setmanalment, per tant, és troben en una freqüencia mtijana i han de tenir una puntuació acceptable.	20	**7**	1,4
Credibilitat i reputació	Aquesta eina mitjançant una cerca ràpida sembla tenir una reputació bona. Cal tenir en compte que té darrera l'equip de Kantar, fet que incrementa la seva reputació.	10	**9**	0,9
Completesa	En aquestes dades no existeixen N/As. Les facturacions resulten rellevants per a tots els clients i tot i tenir menys rellevància en la facturació final, de cara al control actiu, és una de les fonts més usades i per tant, amb més correcions.	20	**10**	2
Representació coherent	Els canvis en aquest datasource són gairebé nuls. A més a més, quan és donen, és manté durant un temps provisional la versió antiga per a facilitar el canvi del sistema.	10	**7**	0,7
Llibertat d'errors	Els errors son existents però al existir correccions diàries, és redueixen al llarg del temps. Aixi doncs no és pot calcular l'index de la manera indicada perquè el nombre d'errors no és estàtic sinò decremental. Per tant s'assigna de manera subjectiva.	15	**9**	1,35
Rellevància i valor afegit	La font aporta el valor de la desagregació per target, però no aporta nova informació respecte d'InfoAdex.	10	**6**	0,6
Puntualitat	Les dades són actualitzades setmanalment, per tant, no són gaire freqüents i però sempre arriben dins del plaç.	15	**9**	1,35
TOTAL			**8,30**	

Figura 8.6: CBR per Kantar TNS [iope]

Observem a la taula 8.8 que la puntuació final és de 8.3. Aparentment, ens trobem davant d'una font amb una qualitat alta que pot ser usable per al cas d'estudi. Trobem també, que tots els diferents subíndexs resulten tenir una puntuació acceptable menys en el cas de valor; aquest fet és degut al fet que, al final, l'aportació de dades no és diferent de la feta per Info Adex, sinó que tan sols afegeix el valor de les agrupacions i freqüència d'aquestes. Així doncs s'ha decidit remarcar aquest fet en el CBR per tal de tenir present la possible futura millora d'aquest sistema per a oferir més desagregació passant així Info Adex

a ser una font sense aportacions rellevants i substituïble per Kantar TNS.

Una vegada donada una qualificació de la qualitat, passem a generar l'últim apartat d'aquesta anàlisi de la font de dades, és a dir, procedim a dissenyar l'UML de l'estructura de classes:

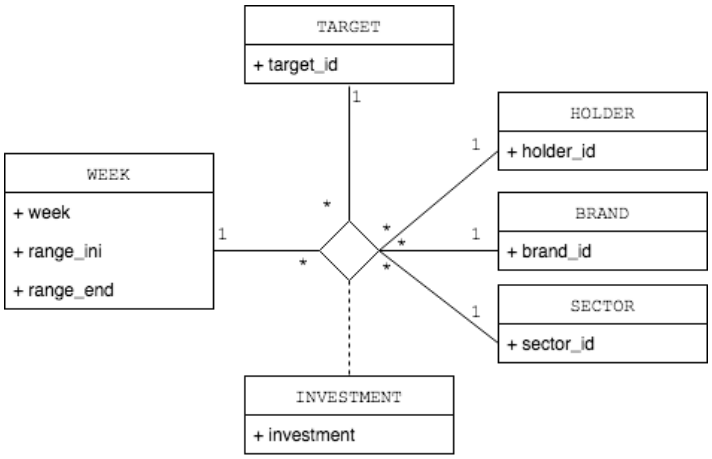


Figura 8.7: Diagrama de classes UML per a Kantar TNS [iope]

Finalment, i acabant amb aquesta font de dades, s'ha realitzat l'UML de la figura 8.7. En aquesta, podem veure que és un UML en el que es creuen classes prèviament definides per altres fonts de dades per tal d'acabar generant una classe on s'acaben guardant les dades d'inversió setmanal. L'única classe no existent és la classe de l'objecte Week, que és simplement una classe que manté el nombre de setmana de l'any i el rang de dates que la defineixen segons Kantar TNS. S'haurà de revisar aquesta estructura de dades en un futur quan s'intentin integrar les diferents fonts per tractar de reduir-ne el nombre de relacions per generar la classe d'Investment, però actualment l'estructura de classes és correcte. Un altre fet a plantejar-se és la possibilitat d'emmagatzemar les dades a escala diària a partir del càlcul de la mitjana diària d'inversió, permetent així guardar-ho amb una major granularitat.

Un cop presentat el disseny UML d'aquesta font de dades, donem per conclòs l'estudi d'aquesta i l'estudi de les fonts globals i passem a estudiar les fonts específiques i extres.

8.3.4 Específiques - Client

Una vegada explicades les fonts estàtiques, és a dir, les globals, cal passar a realitzar l'anàlisi de les fonts de client. Primerament, val la pena realitzar una explicació del perquè aquestes fonts no s'estudiaran tan detall com les globals.

Tot i resultar quelcom poc òptim, ens trobem en una situació en la qual existeixen grans problemes amb les dades que es reben per part dels clients a causa del format de les mateixes i les especificitats que cada una expressen. Resulta lògic que les dades enviades per cada client no continguin unes mateixes dades, ja que, cada un dels clients realitza les campanyes en sectors diferents i, a més a més, busquen resultats diferents, per exemple: un client pot estar interessat en incrementar el nombre de ventes que realitza mentre que un altre pot estar interessat en el nombre de registres web i trucades que rep (KPI). Així doncs, veiem que no es podrà dissenyar una estructura de dades que emmagatzemi la informació de totes les fonts de client, ja que, no tenen cap relació entre elles.

Fins aquest punt, tot sembla gestionable i treballable però els problemes comencen a partir d'aquí. En la realitat no tan sols succeïx que cada client envia dades diferent (fet totalment justificable) sinó que sovint, un mateix client realitza entregues de dades diferents entre elles i freqüentment, en formats totalment canviants depenent de la persona de contacte que envia les dades. Aquests formats, no són tan sols a nivell fitxer sinó que també poden arribar a ser diferències estructurals en les dades, fomentant l'aparició de nous camps, la desaparició de camps existents o fins i tot la modificació de les dades enviades. S'ha remarcat molt aquest fet per part dels diferents components de l'equip i s'ha citat un cas en específic que permet entendre la dificultat existent en l'extracció d'aquestes dades que s'expressa a continuació:

> **Citació 10** *"Treballem amb el client XXX des de fa 3 anys i en aquest temps hem tingut contacte amb 2 interlocutors interns que gestionen les dades i ens les envien mensualment. El fet és que, en un inici teníem contacte amb el/la XXXXX que ens passava la informació fins que va deixar de treballar per l'empresa XXXX i se'ns va assignar un altre interlocutor. Aquest nou interlocutor/a va ser i segueix sent la persona que ens passa les dades per als estudis de l'empresa XXXX. El principal problema és que, el nou interlocutor no tan sols no ens ha passat les dades amb el mateix format, fet que suposa tan sols un canvi de tractament, Simó que, a més a més, ens passa menys dades. Això se li ha notificat i no té constància de l'existència de les dades que li demanem tot i que, sabem que existeixen perquè les hem tingut i hem treballat amb elles quan teníem contacte amb l'altre interlocutor/a"*

Aquesta anècdota ens permet entendre millor el cas d'estudi. La situació real és que, tot i que els clients tenen la necessitat de rebre una justificació numèrica de la presa de decisions, encara no entén el funcionament del servei i no coneixen les bases de qualsevol anàlisi, és a dir, les dades. No resulta estrany rebre dades no coherents dins d'un mateix client o rebre les dades molt fora de termini. Un altre fet a destacar és que, a part de totes les característiques explicades, sovint les dades enviades no són completes o existeixen problemes en aquestes. Tot i això, cal valorar i tenir en compte també que existeixen clients que tenen un servei de col·lecció de dades molt correcta i que no posen tots els impediments prèviament explicats.

Ens trobem per tant davant d'una situació en la qual existeixen un nombre molt elevat de fonts de dades amb una representació i un contingut molt variant. Per tant, s'ha decidit no realitzar una anàlisi exhaustiva de cada una de les fonts sinó realitzar aquest estudi genèric, ja que, existeixen moltes probabilitats que durant el temps en el qual es realitza aquest treball, algunes d'aquestes fonts canviïn el seu format i contingut, deixant desfasada tota la feina realitzada. Això però, serà un fet que s'haurà de tractar de cara al disseny, forçant la solució a ser molt canviable i no dependent d'aquestes fonts. És a dir, aquestes fonts

no poden ser la base del sistema, sinó que hauran de ser un afegit a les dades globals per tal de facilitar el manteniment i el canvi flexible de l'emmagatzematge de les dades, que estem forçats a oferir per culpa de la implicació donada pel client. Cal remarcar també que el fet d'integrar aquestes dades amb dades que continguin identificadors de marca o identificadors de dates no serà difícil, ja que, totes les fonts de dades d'aquest tipus estan agregades per Marca, Model o Data fet que, a primera vista, fa que totes aquestes dades siguin vinculables amb totes les dades globals prèviament tractades.

Així doncs, d'aquestes dades tan sols s'oferirà un estudi de la seva qualitat basat en l'índex CBR, ja que, al no concretar no podrem arribar a especificar ni la taula de meta dades ni el disseny de l'estructura de les dades existents. Presentem doncs la següent taula (Taula 8.8) de puntuació dels subíndexs del CBR:

Subindex(BCI)	Justificació	Pes (W)	Avaluació	Aportació
Quantitat	Totalment dependent del client, però solen ser dades amb quantitats adequades a l'estudi a realitzar. Per tant rebran una punutació alta en aquest àmbit	20	**8,5**	1,7
Credibilitat i reputació	Degut a la gran quantitat d'errors que presenten i la validessa únicament interna de les dades, tenen una credibilitat i una reputació baixa.	10	**5**	0,5
Completesa	Existeixen aproximadament un 7% de N/As en les dades rebudes l'últim més. Tenien en compte com a N/As també dades de dies no enviats que haurien d'haver estat presents pels estudis.	20	**6,5**	1,3
Representació coherent	Els canvis en els datasources son molt freqüents i totalment arbitraris. No donen temps per a l'adaptació.	10	**2**	0,2
Llibertat d'errors	Existeixen errors en les dades i sovint no es poden solucionar. Resulta impossible identificar-los i es freqüent detectar-los un cop acabats els anàlisis degut a que els resultats no son els esperats.	15	**3**	0,45
Rellevància i valor afegit	La font aporta molt valor afegit, és l'única conexió directe amb els KPIs gestionats pel client	10	**10**	1
Puntualitat	Les dades són entregades sovint fora de plaç i resulta necessari reclamar-les per rebre-les.	15	**5**	0,75
TOTAL			**5,90**	

Figura 8.8: CBR per a fonts de dades específiques [client]

S'ha acabat assignant una puntuació total de 5.9 de qualitat a aquestes dades a causa de les seves clares mancances en credibilitat, completesa, coherència, llibertat d'errors i puntualitat. Aquesta qualitat queda amb una nota superior a 5, ja que el valor que aporten les dades específiques a les anàlisis és suficientment representatiu com per usar-les. Així doncs, considerem que aquestes dades presenten una qualitat molt baixa però que han de ser usades degut a la seva importància dins de les anàlisis. Probablement el camí a seguir per millorar aquesta qualitat passa per fer entendre al client la importància de les dades en el servei ofert i en el futur de les empreses per tal que, internament, es fomenti una millora de les dades.

Concloem doncs amb les fonts específiques, sabent que, s'hauran d'usar encara que la seva qualitat no sigui del tot desitjable.

8.3.5 Específiques - Extres

Acabem l'estudi de les diferents fonts de dades específiques parlant de les fonts extres de dades. Les fonts extres de dades són dades externes a tot el món publicitari que es poden prendre per tal d'usar-se com a factor de certs estudis analítics. Per exemple, és sabut que com més temperatura, menys audiència. En aquest cas, la temperatura seria una variable extra, doncs teòricament no hauria de tenir cap relació amb la publicitat però es dóna el cas que sí que té un impacte en les persones que consumeixen aquesta publicitat i, per tant, acaben influint en l'estudi. Així doncs, les dades extra, són dades arbitràries que se seleccionen en dependència de l'anàlisi a realitzar. Per tant, són dades que no necessàriament s'hauran d'emmagatzemar. Un bon criteri per a decidir si una dada extra s'ha d'emmagatzemar podria ser les vegades que s'ha hagut d'usar aquella dada en estudis, guardant així únicament un històric de les dades rellevants i no de totes les dades usades.

Concloem doncs que aquestes dades no seran emmagatzemades a no ser que sigui un cas molt reiteratiu o un cas que aporti una utilitat real. És important remarcar que, tot i no emmagatzemar-les, es poden afegir sobre els conjunts de dades que s'obtinguin de les consultes realitzades afegint un pas més de preprocessing.

9

Estudi de Context: Anàlisi del hardware

Passem ara a realitzar una anàlisi del hardware existent en l'empresa. Cal dir que de cara a aquest punt no es podrà arribar a profunditzar gaire com que la informació que s'ha donat no és completa, ja que, segons sembla, tot està controlat per una empresa externa i no existeix un coneixement clar de l'estructura que hi ha muntada. Presentem el següent esquema que suposadament presenta la tipologia de la xarxa:

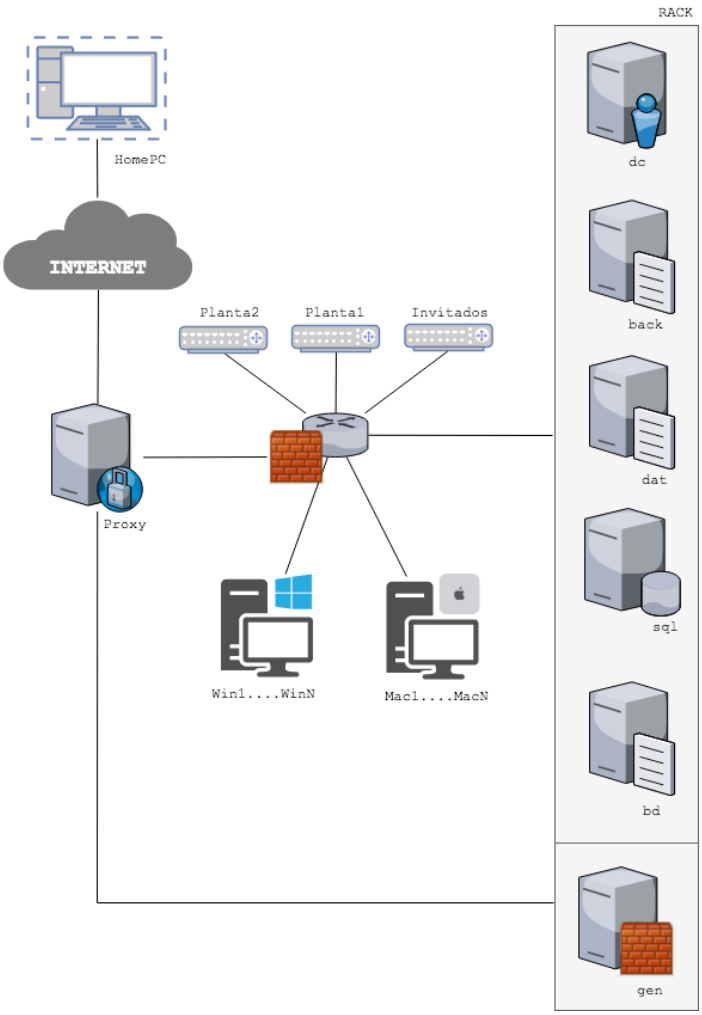


Figura 9.1: Topologia de la xarxa

Inicialment veiem que el servidor usat és un rack que està emplaçat físicament en l'empresa, tot i que, com ja s'ha dit, es gestiona externament. Aquest servidor ("rack") conté 6 servidors virtuals:

- **dc** *(directory sever)* És un servidor que manté els directoris tant de correu com privats dels usuaris que s'usen a IKI Media Communications S.L.
- **back** *(file server)* És un servidor que manté les còpies de seguretat i backups.
- **data** *(file server)* És el servidor que manté els fitxers compartits entre els treballadors.
- **sql** *(sql server)* És el servidor preparat per a mantenir una base de dades relacional.
- **bd** *(file server)* És el servidor que manté les aplicacions i eines de l'empresa. També manté les lectures de les fonts de dades (scripts).
- **gen1** *(desktop server)* Aquest és un servidor que s'usa per a connectar-se des de l'exterior de l'empresa i que ofereix un sistema d'escriptori remot per poder treballar des de l'exterior.

Observem també que els ordinadors que es troben dins l'empresa es connecten als primers 5 servidors virtuals mitjançant el router o els HUBS Planta1 o Planta2. Pel que fa al HUB Invitados, tot i que permet l'accés a internet, no permet mai accés al servidor.

Aquests ordinadors tenen dos sistemes operatius diferents que són: Windows 10 o Mac OSX Sierra. Existeixen aproximadament uns 35 ordinadors en l'empresa (no es concreta el nombre exacte ja que, molt probablement el nombre augmentarà durant els mesos de treball).

Pel que fa a les connexions externes de l'empresa, aquestes passen directament al servidor gen1 que posteriorment es connecta als altres mitjançant el router al que està connectat. Així doncs, ningú es pot connectar externament als 5 primers servidors sense entrar al servidor gen1.

Per acabar de tancar les dades de la figura 9.1 cal dir que, totes les connexions realitzades pel router estan protegides mitjançant un sistema de proxy que redirigeix les connexions per evitar que és pugui realitzar cap tipus de seguiment dels senyals.

Una vegada explicada la informació bàsica recollida de la topologia de la xarxa, passem a explicar els punts més rellevants i que més ens podran servir per a la proposta de solució. Així doncs, creiem que l'únic punt que aportarà valor, serà estudiar les propietats del servidor SQL i les possibilitats d'ampliació del mateix, ja que, al final serà el servidor sobre el qual haurem de muntar la implementació. Exposem doncs, una imatge amb les especificacions del servidor:

Figura 9.2: Especificació Servidor SQL

Finalment, i per concloure, sembla rellevant anotar que la connexió a aquest servidor serà basada en la windows authetication guardada en el servidor virtual DC.

10

Estudi de Context: Anàlisi dels softwares

En aquest apartat es pretén realitzar una anàlisi completa dels softwares utilitzats. Aquest estudi ens haurà de permetre entendre quines són les eines usades i el motiu. Facilitant així, que la nostra solució pugui integrar-se en els softwares en funcionament actual; fet que, teòricament, hauria d'ajudar a reduir la resistència al canvi de sistema de funcionament.

Amb l'objectiu de realitzar una anàlisi de cada software, el primer pas serà realitzar una identificació dels diferents softwares en ús. Per tal de fer-ho, es mantindran un seguit de reunions amb els diferents equips que permetran saber el software existent i l'ús que se'n dóna. D'aquestes reunions s'ha extret la següent llista de softwares a estudiar:

- Microsoft Office, Excel.
- R Studio.
- Tableau desktop.
- Instar Analytics.
- InfoIO.

Una vegada determinats els softwares rellevants en ús, comencem a realitzar l'anàlisi individual de cada un d'aquests que es basarà en una descripció de la funcionalitat del software, una explicació de les tasques que es porten a terme dins d'IKI Media Communications S.L amb el software en qüestió i la valoració de l'ús d'aquestes. També resultarà important veure si els diferents softwares tenen necessitat d'interactuar amb les dades que es busca posar en la base de dades i, en cas afirmatiu, comprovar si el software és integrable i compatible amb les bases de dades SQL o MSSQL.

Figura 10.1: Icones del software

10.1 Microsoft Office, Excel

MS Excel és una aplicació o programa que ha estat desenvolupat per Microsoft i que forma part d'un paquet de programes per a l'oficina anomenat Microsoft Office. MS Excel, es basa en un full de càlcul que et permet crear i manipular taules de dades, gràfics, bases de dades, etc. L'interessant de MS Excel és que pots desenvolupar veritables miniaplicacions avançades que es convertiran en potents eines de treball. Excel pot automatitzar gran part del treball gràcies al fet de donar un sistema de treball mitjançant la interacció amb un front-end per a un usuari bàsic i a la vegada oferir un sistema de treball més tècnic basat en la codificació de tasques mitjançant llenguatges de programació que automatitzin les tasques a realitzar i comprovin que els processos s'estan realitzant de manera correcta. Per tant, MS Excel és un software molt adient per una oficina multidisciplinar on, un conjunt de persones poden fer servir les eines bàsiques d'Excel i d'altres poden dedicar-se a generar macros o eines per a facilitar les feines que realitzen els usuaris bàsics del sistema.

Excel a més, ofereix també facilitats de cara als desenvolupadors, ja que, a part de donar un front-end ja creat, dóna un entorn plenament optimitzat que permet realitzar tractament de dades molt més ràpid que el realitzat tant manualment com mitjançant programes externs no optimitzats. Per tant, aparentment, MS Excel resulta ser un software que dóna moltes opcions de cara al tractament i l'ús de les dades.

En el cas específic de l'empresa, Excel s'usa bàsicament per 3 tasques diferents: la primera és el pre-processing o tractament de les dades que es fa a partir d'Excel i un conjunt de macros que a l'executar manualment, ajuden a l'usuari a refactoritzar tots els paràmetres per tal d'acabar tractant les dades fins a arribar a un estat de correctesa adequat; la segona tasca per la qual s'usa és per a la realització d'anàlisis. Finalment la tercera, que molt probablement és l'única que no encaixa amb la definició de funcionament donada prèviament, és la utilització d'Excel com a històric de les dades.

El fet d'usar Excel com a mètode d'emmagatzemant porta molts problemes, ja que, el software no està pensat per a realitzar aquesta funcionalitat. Comencem així plantejant un dels primers problemes que és que, MS Excel, té un límit de 1.048.576 files i de 16.384 columnes; per tant, suposant que llegim uns 3 anys d'un sol sector de Kantar Media, ja es necessita tenir més d'un fitxer per a guardar la base de dades. A més a més, a aquest factor cal sumar-li que quan un fitxer MS Excel té guardades tantes files, els ordinadors d'oficina no solen carregar-lo ràpidament i solen patir errors que sovint suposen el tancament d'Excel i la pèrdua del treball realitzat. Per tant, podem arribar a dir que aproximadament la capacitat real d'un Excel és d'unes 600.000 files com a molt donats els processadors amb els quals s'està treballant. A tot això també cal sumar-li que, MS Excel no permet generar vincles entre les diferents columnes de les taules i que, per tant, no poden existir claus foranes entre taules.

Pel que fa a les capacitats del programa quant a treball amb bases de dades externes veiem que, Excel permet tant vincles amb fulls de càlcul externs com amb bases de dades SQL amb ODBC. Per tant, és un software que ha de permetre la integració amb el sistema de base de dades que s'ha de muntar. Cal recordar que per restricció haurà de ser un sistema relacional basat en SQL o MMSQL.

Veiem doncs que MS Excel és un software que s'adequa a les necessitats del cas d'estudi però que s'ha de deixar d'usar com a base de dades, ja que, el software no està preparat per suportar-ho ni per realitzar-ho.

10.2 R Studio

RStudio és un entorn de desenvolupament integrat (IDE) basat en el llenguatge de programació R dedicat a la computació estadística i la graficació. Inclou una consola, un editor de scripts que permet executar el codi i eines que permeten la traçabilitat, la depuració i la gestió de l'entorn de treball. És un software cross-platform, és a dir, que es pot utilitzar tant en Windows, com en Linux, com en Mac. És programari lliure en el qual es pot contribuir fet que fa que existeixi molt contingut de la comunitat o d'agrupacions de desenvolupadors. Aquest software ofereix una funcionalitat de descàrrega i utilització de llibreries externes instal·lables via la consola de comandes. Aquest fet la fa encara més interessant, ja que, es poden trobar llibreries que permetin la realització de tasques molt més elaborades que les que el mateix R permet sense necessitat de ser un expert. És el cas de la llibreria usada per a la graficació ggplot i ggplot2 (que són llibreries que permeten la creació de gràfiques més elegants visualment i que afegeixen funcionalitats de cara a la creació de gràfiques representatives de dades).

En l'entorn laboral del departament es treballa amb RStudio (R 3.4.1) per a la realització dels diferents estudis de modelatge i anàlisis descriptius. En el seu moment es va decidir usar aquest software (IDE), ja que, resultava molt difícil per a una persona amb coneixements bàsics de programació treballar en l'entorn ofert per la consola amb R. A més a més, d'oferir els trets bàsics d'un IDE, és a dir, ressaltat de sintaxi, autocompletat, sagnat intel·ligent, etc. Dóna d'altres funcionalitats com ara l'execució parcial de blocs de codi i la capacitat de declaració de funcions amb front-end integrat. Aquest fet, permet a les persones més tècniques implementar scripts en forma de funció que posteriorment poden usar els usuaris més bàsics per tal de realitzar accions complexes d'anàlisi. És a dir, dóna l'opció de dividir la feina de programació. Els tècnics de programació (com l'analista programador) poden generar codi que doni com a resultat X i l'estadístic, sense haver d'entendre com es fa X pot executar-ho i obtenir X sense necessitat d'implementar-ho.

Un altre punt a tenir en compte és la quantitat de documentació que existeix del software i del llenguatge, arribant a oferir un sistema de documentació i suport integrat permetent als usuaris revisar la documentació de funcions. Aquest fet, en el cas d'estudi, aporta un valor afegit, ja que, resulta d'ajuda per a una persona amb coneixements bàsics de programació tenir disponible un sistema d'ajuda per, en cas de no conèixer les possibilitats ofertes per una variable o llibreria, poder consultar sense necessitat de sortir de l'IDE.

Així doncs, com que per a l'ús d'aquest sistema també serà necessària la introducció de les dades, s'ha d'assegurar la seva possible integració amb la base de dades. Aparentment, existeixen llibreries que permeten gestionar la connexió de R amb les diferents bases de dades existents, com és el cas de RODBC (un paquet que gestiona les connexions ODBC amb bases de dades). Fins i tot R, té un sistema integrat de connexió a bases de dades a les quals tan sols proporcionant el driver necessari per a la connexió i les dades de connexió (ip, port, usuari...) facilita la gestió de la connexió. Inicialment doncs, s'haurà de provar la versió integrada de R, ja que, en cas de funcionar correctament, es podrà prescindir de la dependència generada per l'ús d'una llibreria i els problemes que poden suposar les actualitzacions sobre les llibreries externes. Per tal de provar-ho però, s'haurà de descarregar els drivers adequats per a la base de dades triada, decisió que serà presa més endavant (molt probablement ODBC, al ser compartida per Microsoft Office).

Veiem doncs que RStudio és un software que s'adequa a les necessitats d'anàlisi de l'equip de Data Science i que té mètodes vàlids de connexió tant amb bases de dades SQL com amb bases de dades MSSQL

10.3 Tableau desktop

Tableau Desktop Professional és un programari de visualització i anàlisi de dades que a diferència d'altres (en els que s'ha de programar) a través d'una interfície d'arrossegar i deixar anar, permet que els usuaris puguin crear i compartir visualitzacions interactives i dashboards. L'ajuda principal que ofereix Tableau sobre altres softwares és que al fer tan simple la creació de gràfics, redueix el temps de creació d'aquests. A més a més, permet introduir dades de més d'una font de dades (sempre que mantinguin una mateixa estructura per a la realització del encreuament) i vincular-les per tal de poder fer representacions conjuntes de dades extretes de diferents fonts.

Com ja s'ha dit abans, les visualitzacions o graficacions generades són molt interactives i és possible la creació d'aquestes de manera molt senzilla. Existeixen múltiples possibilitats de gràfic, com ara: de barres, circulars, mapes de calor, gràfics d'àrea terrestre, d'entre una infinitat més d'opcions. Recordem també que aquestes gràfiques seran modificables a temps real per l'usuari que les vegi amb un mètode simple de 'drag and drop' a l'estil de les taules dinàmiques d'Excel i aquest fet, permetrà que en una mateixa visualització puguem veure diferents dades de manera simple un cop generada l'estructura gràfica inicial.

A part, i com a una de les coses més importants que ofereix, existeix la possibilitat de compartir amb persones alienes al projecte els diferents gràfics generats, donant així opció a què es pugui compartir via pdf, correu o web els estudis realitzats a mode de dashboard. El fet de compartir-lo via web ofereix una oportunitat més que s la de donar als clients el contingut treballat en temps real. La interacció amb les dades, és a dir, els clients també tindran accés a interaccionar amb les gràfiques per veure diferents visualitzacions a partir de les dades existents.

Pel que fa a IKI Media Communications S.L. , aquest software/plataforma s'usa amb l'objectiu de generar dashboards actualitzats diàriament que permetin als clients visualitzar gairebé a temps real (a 1 dia vista) l'impacte i l'evolució de la campanya que s'està portant a terme. Per tal de donar aquest servei, es genera a partir de Tableau Desktop un dashboard que mantindrà totes les dades que es volen usar per ensenyar al client i es prepara una estructura de dashboard que mostri visualitzacions útils de les dades introduïdes. Posteriorment, es publica tot a Tableau Web, que serà accessible tan sols a través de la compta del client i aquest tindrà accés a la lectura sobre els dashboards generats, és a dir, que podrà veure les dades i interaccionar amb les gràfiques però mai podrà realitzar canvis sobre el treball creat ni accedir directament a les fonts de dades o a les dades planes. Per tant, amb aquest software el que s'aconsegueix és acostar de manera molt senzilla i totalment separada de la tecnologia que hi ha darrere, les dades als clients, donant-los accés a les visualitzacions de les gràfiques però mai a les dades o a les fonts.

Aquest programa també haurà de poder tenir una integració completa amb la base de dades, ja que, és totalment dependent de tenir les fonts de dades per tal de poder-ne facilitar la visualització. Es comprova que existeix la possibilitat de vincular aquest software amb bases de dades SQL i MMSQL, fulls de càlcul (MS Excel) o aplicacions al núvol (com Google Analytics i Salesforce). Aquest fet ens assegura que serà una plataforma plenament integrable amb el sistema a dissenyar.

10.4 Instar Analytics

Instar Analytics és un programa d'explotació de la base de dades KM prèviament explicada. Així doncs, aquest software funciona com a eina d'extracció de dades de la font de dades KM proposant-nos un sistema visual de realització de queries que ens permeten recollir les dades que volem filtrades pels diferents camps d'agrupació prèviament explicats en el punt d'anàlisi de les fonts de dades i basant les mètriques en els targets i àmbits seleccionats. Aquest sistema d'interacció i recollida de dades tant user-friendly acaba permetent que qualsevol persona amb una mica d'experiència pugui executar queries sobre les dades de Kantar Media i aconseguir el conjunt de dades que desitja per a la realització de les seves tasques. A més a més, la mateixa plataforma ens ofereix un API per a l'explotació de les dades, fet que dóna l'oportunitat a connectar les eines pròpies amb les dades oferides per Kantar Media.

A l'empresa IKI Media Communications S.L l'ús que es realitza d'aquesta eina és bàsicament el descrit anteriorment com a funcionalitat real. És a dir, que s'usa per a l'extracció de les dades amb les mètriques i filtres desitjats per a l'anàlisi. El fet que la interfície sigui molt user-friendly permet que els treballadors de tots els àmbits, siguin capaços de tractar les extraccions de dades i generar les seves pròpies plantilles de queries per tal d'aconseguir els datasets que necessiten per a la realització de la tasca que tenen assignada.

A més de les queries que poden generar els diferents usuaris interaccionant amb el programa, el software permet guardar i establir plantilles genèriques per a l'equip. Aquest fet, dóna la possibilitat d'oferir formats de query que extreuen un format i tipus de dades estàndard, és a dir, la funcionalitat de les plantilles acaba oferint l'opció de donar als treballadors una query preprogramada que traurà un dataset en el que podran trobar les dades que necessiten expressades en un format estàndard que les macros de tractament de MS Excel poden arreglar, ja que, mantenen una mateixa estructura i un mateix contingut. Posteriorment, i un cop tractades les dades per les macros, els treballadors tan sols s'han de dedicar a seleccionar el subconjunt de dades del dataset tractat per les macros que necessiten per a la tasca.

Actualment doncs, el sistema més genèric i automàtic de tractament de les dades està basat en el flux explicat, és a dir, en què l'usuari demani un conjunt de dades mitjançant una plantilla al software i que posteriorment utilitzi l'execució del tractament automàtic implementat en macros amb la resposta donada aprofitant que està estandarditzada i sempre té una mateixa estructura. Com a resultat es genera un fitxer de MS Excel preparat per a treballar.

Un fet que s'ha descobert per part meva durant la realització d'aquest estudi és l'existencia d'un API que podria arribar a permetre executar les queries des d'una les aplicacions de l'empresa o des de scripts automàticament executades que preparessin les dades abans que els diferents treballadors arribessin al lloc de treball i carreguessin la base de dades directament sense interaccions per part de l'usuari.

Pel que fa a aquest software no té cap necessitat d'integrar-se amb la base de dades, ja que, no necessita dades sinó que n'és el proveïdor. Tan sols fa falta que sigui capaç d'exportar-ne, fet que, com s'ha vist anteriorment, compleix.

10.5 Info IO

Info IO és un programa d'explotació de la base de dades INV anteriorment explicada. Per tant, bàsicament és una aplicació que permetrà a un usuari bàsic realitzar les consultes sobre la font de dades INV mitjançant una aplicació visual que pretén facilitar-les. Mitjançant aquest software per tant, es busca donar opció de treballar a qualsevol dels stakeholders amb les dades d'inversió sense necessitat de ser un especialista o tenir coneixements en SQL, APIs o llenguatges específics. A més a més, també dóna l'oportunitat a l'usuari de decidir el grau d'agrupació que vol en les dades, facilitant així que cada un dels stakeholders aconsegueixi un conjunt de dades vàlides i adaptades a les seves necessitats. Així doncs, ens permetrà segmentar i agrupar les dades en dependència a múltiples variables com: marca, model, anunciant, mitjà, suport, forma d'anunci, agència de mitjans, àmbit, província, etc.

Pel que fa a IKI Media Communications S.L., aquest software/plataforma s'utilitZA amb l'objectiu d'aconseguir les dades d'inversió per sectors. Així doncs, aquest software dóna accés a les persones de l'equip de Data Science i Contabilitat a un registre complet de les inversions pròpies i dels diferents sectors treballats. Ens trobem doncs, davant d'un software que no és usat per gran part de l'empresa, ja que, els estudis de competència són directament portats per l'equip de Data Science i la gestió d'inversió i control de costos és portat per part del departament de comptabilitat (que també té accés a altres softwares que permeten un control més exhaustiu però que no seran tractats, ja que, no s'usen en el scope del nostre projecte).

Un altre punt a tenir en compte és que, a diferència de les extraccions realitzades en Instar Analytics, en aquest cas no existeixen macros de tractament de les dades que s'extreuen, ja que, el format estàndard d'aquestes és suficientment entenedor per a ser tractades manualment i usades directament. A més a més, aquest fet permet també que, a diferència del cas d'Instar Analytics, els usuaris puguin realitzar consultes pròpies i no tan sols les consultes estàndard pregenerades que la macro sap tractar. Aquest fet també va molt lligat a què les persones que usen aquest software són persones amb perfils més tècnics i que, per tant, poden autogestionar les dades i això no els suposa un esforç o temps extra rellevant.

Finalment, cal ressaltar la importància d'un dels serveis extres que ofereix aquest software que és independent a l'extracció de dades. Aquest software ofereix un informe predissenyat amb les diferents explicacions de l'evolució del sector publicitari a nivell general que permet veure les tendències a final de cada trimestre. Aquests informes a diferència del software en si, són distribuïts per tota l'empresa i es porta un control perquè cada un dels treballadors realitzi una lectura del document. Aquest fet resulta rellevant, ja que, dóna a tots els treballadors una formació constant de l'evolució del mercat, fet que, a llarg termini, permetrà que les diferents decisions realitzades a l'empresa puguin encaminar-se cap a les tendències i mantenir sempre el servei ofert a l'alçada dels altres serveis existents en el sector publicitari. Un cas pràctic en el qual s'ha vist reflectida aquesta importància és en el cas de la publicitat online, que, vista la tendència del sector a caure en favor de l'augment d'inversió en aquest mitjà, s'ha decidit augmentar l'oferta en aquest tipus de publicitat i també augmentar-ne la importància dins de l'empresa.

Pel que fa a aquest software no té cap necessitat d'integrar-se amb la base de dades, ja que, no necessita dades sinó que n'és el proveïdor. Tan sols fa falta que sigui capaç d'exportar-ne, cosa que com s'ha explicat abans, es compleix.

10.6 Conclusió softwares

Concloem doncs amb l'estudi dels diferents softwares usats en l'àmbit de treball. Cal tenir en compte que, dins de l'empresa, s'usen més softwares que no han estat estudiats de cara a aquest informe perquè no tenen una influència sobre el projecte que s'està portant a terme.

Podem determinar que els softwares usats són vàlids per l'ús que se'n dóna i que tots hauran de ser tinguts en compte de cara al disseny d'una solució. Especialment en el cas de Microsoft Office Excel, R Studio i Tableau Desktop, ja que són softwares que s'hauran de poder connectar amb la base de dades per tal de poder obtenir els conjunts de dades amb els que treballar. Tot i això, també caldrà adaptar el disseny a les dues eines d'extracció per fer-lo més eficient i s'hauran de tenir en compte els formats de dades que proporcionen de cara a la introducció de les dades a la base de dades centralitzada.

11

Estudi de Context: Anàlisi dels stakeholders

Procedim ara a realitzar un estudi exhaustiu dels stakeholders anomenats anteriorment en l'estudi de context introductori. Per tal de fer-ho, volem començar justificant la gràfica 2.2. presentada que tornem a exposar a continuació per tal de facilitar-ne la lectura i la posterior comprensió de l'explicació que se'n donarà.

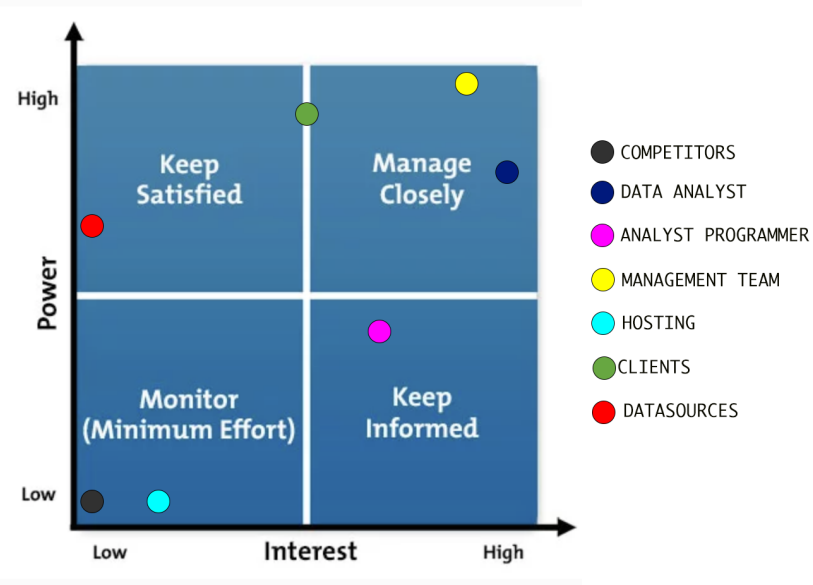


Figura 11.1: Reixeta per a la prioritització dels stakeholders: Poder vs. Interés

Aquesta gràfica té l'objectiu de donar una idea del seguiment que s'ha de fer de cada stakeholder i la posició de rellevància dins del projecte que mereix cada un d'ells. Com molt bé indica, l'eix d'abscisses representa la importància de mantenir el stakeholder informat, mentre que l'eix d'ordenades representa la capacitat que té aquest de tenir influència sobre el resultat final del projecte (ja sigui degut a tenir poder de decisió sobre el pressupost o per ser un dels possibles usuaris que definiran les necessitats reals del projecte). La gràfica queda dividida en 4 espais definits en el primer quadrant dels eixos. S'ha decidit dividir en aquests quatre grups tot i que es podria arribar a incrementar el nombre d'espais definits per a una major precisió en l'anàlisi. Tot i això, s'ha decidit deixar-ho en aquests quatre, ja que, a visió personal, resulten més que suficients per a determinar de manera inicial els stakeholders als que seleccionarem per a l'estudi en profunditat.

Així doncs, i com s'ha vist anteriorment en el punt 2, per omplir el gràfic explicat, primerament s'ha hagut de realitzar un estudi superficial dels diferents stakeholders existents en l'àmbit d'estudi amb l'objectiu d'aconseguir una descripció inicial de cada un d'ells i mitjançant aquesta descripció i la informació recopilada s'ha passat a omplir el gràfic definit amb l'objectiu de diferenciar els stakeholders rellevants. Amb l'objectiu de, posteriorment, tractar-los d'estudiar exhaustivament per tal d'entendre'ls millor i per tant, assegurar que el projecte a realitzar serà útil i vàlid per a cadascun d'ells.

D'aquest estudi superficial, el primer que n'hem extret és el llistat de stakeholders mostrats en la figura 2.1. Posteriorment a la selecció inicial de stakeholders, s'ha realitzat una descripció de cada un d'ells que es pot trobar entre les pàgines 5 i 8. Finalment, a partir d'aquesta informació s'ha omplert la gràfica corresponent prèviament explicada i, a continuació, es pretén justificar la col·locació de cada un dels punts per tal d'acabar seleccionant els stakeholders que s'han d'estudiar amb profunditat de manera raonada i no tan sols en base a una gràfica mostrada però no argumentada.

Un cop explicat el procediment, passem a realitzar la justificació de cada un d'ells bastant-nos en l'espai en el qual es troben i la seva possible proximitat a altres espais.

- *Competidors - Mínim Esforç:* En aquest cas s'han situat els competidors amb un interès molt baix, ja que, més que interès demostren desinterès envers el projecte doncs, per a ells, resulta molt més rendible que no acabi funcionant com que, una millora en el nostre servei pot indicar un augment de quota de mercat de IKI Media Communications S.L. Aquest fet repercutiria directament sobre els competidors. Afortunadament, aquests també es troben en una puntuació molt baixa quant a poder de decisió, ja que, no tenen cap manera d'interposar-se en el projecte pel fet que no es troben dins de l'empresa. Així doncs, queden identificats amb la puntuació més baixa en els dos casos quedant així com el stakeholder al que s'ha de donar menys cobertura durant el projecte.

- *Hosting - Mínim Esforç:* Aquests stakeholders són els dedicats a cuidar dels servidors de l'empresa. Donat que, com s'ha explicat abans el sistema hardware és local (punt d'anàlisi del hardware) i que ells tan sols el gestionen via internet, no representen un stakeholder amb molt poder sobre el projecte. De fet, representa més poderós el hardware ja montat que ells mateixos que tan sols s'hauran de dedicar a muntar un servidor virtual sobre el servidor físic ja existent. Per tant, necessitaran que el hardware sigui capaç d'oferir els requisits mínims per establir-hi el potencial demanat. Pel que fa a l'interès, se'ls hi ha donat una puntuació una mica més alta a causa del fet que, a l'haver d'obrir el servidor actual, tenen un interès en què els hi porti el mínim de feina de cara a la creació i el manteniment d'aquest.

- *Data Sources - Mantenir satisfet:* Als proveïdors de dades se'ls ha donat un poder molt elevat, ja que, en cas de canviar el format d'enviament de dades o de tancar el servei, tindrien un efecte molt alt de cara al projecte. A més a més, també se'ls hi dóna un valor molt elevat en poder, ja que, en cas d'incrementar la seva tecnologia o oferir nous sistemes d'explotació de les seves dades, poden afavoir molt al projecte. En canvi, pel que fa a l'interès se'ls ha donat una puntuació baixa, ja que, tot i que ells poden estar interessats en què els clients usin el màxim potencial de les seves dades, de cara al seu negoci el fet que un client (cal remarcar aquest un, parlem d'un cas específic no generalitzat) utilitzi o no adequadament les seves dades no hi té un efecte directe.

- *Analista programador - Mantenir informat:* Pel que fa a l'analista programador del departament, s'ha decidit establir-li una puntuació de poder mitjana tirant cap a baixa, ja que, tot i tenir una veu important dins de l'empresa, al no ser tècnic de bases de dades, la seva capacitat d'influència en el projecte recau més en els consells i pautes que pot arribar a aportar que en l'impacte que pugui tenir en l'equip directiu que confia plenament en el fet que s'ha de realitzar aquest avenç. Quant a interès,

se l'ha situat en una puntuació mitjana tirant cap a alta, ja que, si bé és cert que l'aparició d'aquest sistema reduiria la seva càrrega, també és cert que disminuiria igualment la seva posició de poder dins de l'empresa a causa de l'aparició de sistemes que realitzarien tasques que actualment tan sols ell sap resoldre.

- *Clients - Administrar d'aprop:* Comencem ara amb els que han rebut més puntuació parlant dels clients. Als clients se'ls ha donat una puntuació alta en poder, ja que, en cas de mostrar clarament que augmenta o disminueix la seva necessitat de rebre estudis analítics de les campanyes podrien afectar directament en la realització del projecte. A més a més, si els clients aporten dades més o menys freqüentment, si les aporten en formats semblants o diferents, si revisen les dades que envien, etc. milloren o empitjoren molt el balanç final del projecte. Pel que fa a interès se'ls ha donat una puntuació mitjana, ja que, si és ben cert que els hi interessa rebre aquest servei quan més correctament i freqüentment possible, no els interessa saber la maquinària o el sistema que hi ha a la darrera; així que, de cara als seus interessos que ho fem a mà o en un sistema automàtic no resulta rellevant.

- *Equip directiu - Administrar d'aprop:* Seguim amb l'equip directiu que té un paper important dins del projecte, ja que, són els que finalment decidiran si és rendible la seva implementació. Se'ls ha donat el nivell de poder més alt, perquè qualsevol decisió que prenguin haurà de ser reflectida en el projecte i, a més, en controlen el pressupost. També se'ls ha posat un nivell d'interès bastant elevat, ja que, entre la cúpula directiva existeix una visió de futur molt basada en l'anàlisi de dades i l'exposició per internet de les mateixes per acabar d'acostar el servei al client. Així doncs, tenen un interès real en què el projecte sigui viable i es pugui acabar implementant, ja que, seria la base per a tots els desenvolupaments de dashboards i anàlisi que posteriorment s'oferiria, acostant així els estudis al client de manera automàtica.

- *Estadístic - Administrar d'aprop:* Acabem amb els estadístics de l'equip, els quals tenen una puntuació semblant a la de l'equip directiu per inversa en els eixos, és a dir, tenen el mateix interès que poder i el mateix poder que interès que l'equip directiu. Aquest fet ve atès que, a nivell d'interès, al ser els que acabaran usant el sistema resulta lògic que el seu interès pel mateix sigui el més elevat de tots els prèviament tractats. Pel que fa al poder resulta igual d'evident que, un equip de treballadors acaba tenint un nivell de poder d'afectació sobre el projecte inferior a un equip directiu, tot i que, el poder que tenen els dos stakeholders resulta elevat i molt rellevant de cara a la realització del projecte.

Veiem doncs que els stakeholders que reben una puntuació més elevada en els dos àmbits i per tant els que s'hauran d'estudiar són els stakeholders interns. Cal remarcar que no tots han estat tractats, ja que, s'ha iniciat l'estudi eliminant alguns dels stakeholders interns degut a la seva menor rellevància en el projecte, com ara d'altres treballadors o el desenvolupador del sistema, que al ser la persona que realitza aquest TFG no s'ha trobat rellevant el seu propi estudi ni la seva avaluació per poder arribar a pecar de subjectivitat.

Passem doncs a estudiar els casos de stakeholders que es troben dins d'administració propera amb l'afegit del cas de l'analista programador que s'inclou a la llista per decisió pròpia de cara a no generar una situació de diferenciació entre empleats d'un mateix departament. A més a més, creiem que aquesta inclusió afegirà valor a l'estudi, doncs tot i no ser una persona apoderada de cara al projecte, resulta ser la persona a la qual més dubtes tecnològics se li podran preguntar. Per a la realització d'aquest estudi s'usarà un sistema de taules semblant a l'anterior però més extens.

11.1 Analista Programador

Descripció? L'analista programador actualment és la persona responsable de realitzar l'extracció i el tractament de les dades. A més a més, porta el control sobre qualsevol problema tècnic o informàtic que pugui succeir en l'entorn de treball. A més, és la persona que manté el contacte amb els gestors del servidor i la persona que coneix l'estructura hardware i software muntada a l'empresa. Actualment treballa a jornada completa i té molta experiència en el sector (aproximadament 25 anys). Actualment, utilitza Excel com a repositori de dades i té implementades un cojunt de macros que funcionen per a tots els usuaris de l'empresa a mode de macro. També porta els temes de tableau i d'aplicacions internes com l'ikiplan. Aparentment, és una persona formada a base d'experiències i cursos formatius.

Benefici o perdua? Aquest resulta ser un cas bastant especial, ja que, és una persona que tant rep beneficis com pèrdues, així que procedim a especificar els dos casos. Comencem doncs pel benefici, resulta evident que si es donés el cas que el projecte s'acabés implementant, la seva càrrega diària es reduiria perquè un subconjunt de les tasques que actualment realitza, serien automatitzades i, per tant, apart de realitzar-se més ràpidament i més fàcilment li reduirien la quantitat de feina, cosa que probablement permetria que dediqués més temps a desenvolupar. D'aquest mateix fet, però, surt l'inconvenient que, en el cas que s'implementés una solució, moltes de les tasques que actualment tan sols ell és capaç de realitzar, serien fetes automàticament o accessibles per tota l'empresa. Fet que podria portar a una pèrdua de poder dins l'empresa.

Que vol? Aquest stakeholder vol que es generi un sistema de tractament i emmagatzematge de dades que sigui capaç de reduir la seva càrrega de treball i reduir el risc d'error que sovint li suposa la realització d'hores extra per tal de solucionar problemes existents en el tractament o en la gestió de les dades. A més a més, aquest treballador té la necessitat d'entendre el que s'està realitzant, ja que, en un futur haurà de ser el gestor de la base de dades i controlar el tractament d'aquestes. Un altre punt a tenir en compte és que aquest stakeholder no vol sentir-se desprotegit en el seu lloc de treball i, per tant, caldrà donar-li explicacions detallades de com avança el projecte per tal de mostrar-li que és una eina més de treball que podrà usar i en la que refiar-se. A més a més, també resulta important que rebi tota la documentació generada per tal que en un futur, si s'implementa, sigui més fàcil desenvolupar serveis sobre l'estructura proposada durant aquest TFG.

Com comunicar-nos? Comunicació directa, cara a cara i diària per a facilitar la informació sobre l'avenç del projecte. Dubtes, consells i comunicació via mailing (empresa) per deixar constància de decisions crítiques del sistema en les que estigui involucrat.

Estat actual? Actualment el treballador s'encarrega únicament del tractament de les dades i la solució de problemes informàtics i tècnics existents a causa de la falta de temps. Aquest fet és degut a la quantitat de feina manual que requereix a hores d'ara el tractament de les dades, el sistema d'emmagatzemant i la implementació no modular de les eines de treball que forcen al treballador a canviar gairebé tota la plataforma per tal de modificar l'ús de qualsevol de les eines creades.

Com superar l'estat actual? S'ha d'oferir un software que permeti a l'usuari realitzar de manera més senzilla i automàtica el tractament de les dades. També s'ha d'oferir un sistema d'emmagatzemant òptim que permeti gestionar un històric i realitzar consultes amb diferents formats de sortida que redueixin el temps d'obtenció de les dades tractades. Resulta important donar la documentació del projecte per tal de facilitar una propera integració amb les diferents eines implementades. Un altre punt que podria millorar l'estat actual, seria si s'aconseguís mostrar algun tipus de model modular o una estructura de classes d'un sistema software que permetés al treballador reimplementar o refactoritzar les apps creades per tal de fer-les canviables i mantenibles.

11.2 Clients

Descripció? Els clients són el conjunt d'empreses que contracten els serveis d'IKI Media Communications per a la gestió de les campanyes publicitàries. No tots els clients tenen accés al servei d'anàlisi de dades; tot i que, sovint aquest servei a partir d'un cert pressupost d'inversió es dóna de forma gratuïta. No existeixen clients dins del mateix sector treballant amb l'empresa per evitar competència il·legítima. Els clients que opten al servei de Data Science sovint no tenen tècnics en l'àmbit i creuen fidelment en els resultats donats tot i que existeixen casos de clients amb equips especialitzats en l'estudi amb els que les anàlisis es realitzen tan sols com a revisió de resultats per les dues entitats.

Benefici o pèrdua? Inicialment i si tot segueix igual (condicions de preu dels estudis), el projecte generarà un benefici de cara al client, ja que podrà obtenir estudis analítics més freqüentment. Molt probablement els estudis també es podran realitzar amb una quantitat de factor superior i es podrà dedicar més temps a la comprensió i anàlisi dels resultats fet que, també hauria de beneficiar al client. A més a més, es reduiran els errors en les entregues i en el tractament de les dades, fet que generarà més confiança en els resultats.

Què vol? El client és totalment indiferent de cara al tractament i l'ús que realitzem de les dades (mentre no sigui fraudulent) i, per tant, de cara al projecte és bastant indiferent, ja que, per ell a curt termini no representa un benefici directe. Així i tot, com tot client de tota empresa, el que busca és rebre el millor servei al preu més assequible i per tant, en l'àmbit de l'anàlisi de les dades busca simplement que l'equip li realitzi estudis més freqüents, amb més quantitat de dades que li permetin entendre millor el seu sector i, per tant, invertir de manera correcta per aconseguir un major impacte de cara al seu negoci.

Com comunicar-nos? Teòricament no ens hem de comunicar amb ells però en cas d'haver-ho de fer, ho podem fer via mailing o via trucada telefònica. També es concerten reunions amb els clients a l'empresa en les que es podria arribar a comentar certs punts rellevants pel projecte.

Estat actual? Actualment el sector ha girat envers l'anàlisi de les dades, com ja s'ha explicat en la introducció, i aquest fet ha produït que gairebé tots els clients importants comencin a demanar justificació de les decisions de contractació d'anuncis realitzada mitjançant l'avaluació d'històrics de dades que ens ofereixen i els històrics que contractem. El problema d'aquest creixement tan ràpid en el seu interès és que no hi ha hagut un temps per poder arribar a interioritzar aquest àmbit del Data Science i del món del "Big Data", fet que produeix que sovint no entenguin els resultats de les anàlisis (no s'entenen els nombres negatius ni baixos en depèn de quin entorn tot i que a vegades un valor negatiu o un valor baix pot ser un bon indicatiu) i en especial, no s'entén que les anàlisis estan basades en dades i que, per tant, existeix una dependència directa entre la quantitat de dades i la qualitat d'aquestes+s que ens envien de les seves bases de dades i l'encert o la qualitat de l'anàlisi realitzat.

Com superar l'estat actual? Amb l'objectiu de superar l'estat actual, el primer que s'ha de fer és normalitzar els estudis que es realitzen i exposar-los de manera molt clarificadora per poder arribar a fer entendre al client el valor real que suposen aquestes anàlisis de cara a la planificació de les campanyes i inclús de cara a possibles millores que podrien arribar a aplicar en el seu model de negoci. Una vegada s'hagi aconseguit d'entendre aquest factor resulta de vital importància que es comenci a introduir i integrar en el client el concepte de què aquesta feina no és una feina que pugui fer un servei extern sense la col·laboració del client, ja que, al final el que necessita el departament de Data Science són dades amb les que tractar i tot i que gairebé sempre es reben dades, no sempre es troben dins del termini de temps acordat, no mantén un format estàndard i sovint contenen errors que posteriorment mitjançant les anàlisis es detecten i s'han de solucionar. Per tant, l'objectiu en aquest cas, serà guanyar confiança mitjançant la feina ben feta mentre s'intenta integrar al client perquè faciliti les dades necessàries i entengui el valor afegit que aquest nou

servei li suposa.

11.3 Equip Directiu

Descripció? L'equip directiu és el conjunt de persones que gestionen l'empresa i prenen les decisions. Per tant, també són les persones que tenen poder sobre la decisió en l'àmbit de la inversió. A part, es dóna el cas que sovint les persones que porten els clients formen part de l'equip directiu i per tant, també podríem identificar aquest equip com l'equip dedicat a fer créixer el mercat d'IKI Media Communications S.L.

Benefici o pèrdua? Al ser propietaris de l'empresa en la seva majoria (tot i que alguns minoritaris), qualsevol producte que es generi dins l'empresa els genera un increment de capital, és a dir, la implementació d'un software que permeti el tractament i l'emmagatzemant de les dades els suposa un benefici. A més a més, cal tenir en compte que aquest conjunt de stakeholders tenen una visió diferent de la que podent tenir altres, ja que, per exemple per ells, molt probablement el valor real del projecte a realitzar no és la reducció de càrrega dels treballadors (que òbviament també ho serà, ja que, si ara un treball T suposa una càrrega C i en un futur la mateixa tasca T passa suposar una càrrega C' que cumpleix C ¿ C' voldrà dir que el temps C - C' que l'empleat estarà reduint de T passarà a ser temps útil per a la realització d'una altra tasca i per tant s'incrementarà la capacitat de càrrega total del departament) sinó l'increment de valor del servei ofert que permetrà en un futur donar un valor diferencial per davant de les mitjanes empreses que no ofereixin aquesta millora. A més a més, en el cas d'aconseguir tenir un sistema d'emmagatzemant també creuen que els hi millorarà molt la situació de cara a la implementació de nous softwares (eines) i en la realització d'anàlisis a mode de dashboarding que acostarà encara més el servei al client, donant-li un feedback diari de l'activitat que s'està realitzant. Veiem doncs que resulta de gran importància mecanitzar tot aquest procés de cara a afegir valor al negoci i així produir benefici.

Què vol? El que volen aquest conjunt de stakeholders és que s'acabi implementant un sistema de tractament de dades i d'emmagatzament que els permeti reduir els costos que suposen els riscos d'error i que ofereixin una millora en el servei final. A més a més, també volen que es doni documentació molt detallada del sistema per del de posteriorment poder-lo fer servir de base per a tot el desenvolupament tecnològic relacionat amb l'empresa. Un altre punt important per l'equip directiu és poder arribar a justificar el cobrament pels serveis de Data Science realitzats com a un servei extra diferent a la planificació estàndard.

Com comunicar-nos? La comunicació amb l'equip directiu serà en persona i mitjançant reunions anticipades degut a la seva alta càrrega de feina. Hauran de ser reunions ràpides màxim de 10-15 minuts.

Estat actual? L'estat actual de l'equip directiu depèn totalment de la càrrega de treball. Com s'ha explicat abans estan realitzant la feina de comercials i per tant, la seva feina actual es basa en gestionar i mantenir l'empresa mentre busquen nous clients per a intentar que creixi i augmentar així la facturació i el benefici de la mateixa. Bàsicament s'organitzen entre ells perquè sempre hi hagi un o més directius a l'empresa per solucionar problemes d'execució i mentrestant, els altres es dediquen a fer de comercials i a negociar marges.

Com superar l'estat actual? L'estat actual en el que es troben l'han de superar ells mateixos però una aportació que podem realitzar és la possiblitat d'implementar aquest software que a llarg plaç acabarà afegint un valor significatiu al servei ofert i per tant, teòricament hauria de facilitar la captació de clients. A més a més, la implementació d'aquest software també hauria de reduir el risc d'error reduint conseqÜentment els costos derivats d'errors que es tenen a l'empresa.

11.4 Estadístic

Descripció? Els estadístics de l'empresa són les persones encarregades de, una vegada realitzat el tractament de les dades, recollir-les en un dataset i realitzar les anàlisis pertinents. Tot i això, actualment i com s'explica a l'estudi de context inicial el procés de tractament de dades resulta ser el coll d'ampolla de les anàlisis i, per tant, actualment els estadístics també estan treballant en el tractament de dades per tal de poder entregar les feines requerides dins dels terminis establerts. A més a més, els estadístics són els encarregats d'assistir a les reunions (generalment) de seguiment de les campanyes juntament amb altres persones de diferents equips multidisciplinars per tal d'oferir els seus resultats. És a dir, la justificació de les preses de decisions que es fan durant la campanya.

Benefici o pèrdua? Per a aquest conjunt de stakeholders resulta totalment beneficiós aquest estudi i la realització d'aquest treball, ja que, en cas d'acabar-se implementant podran deixar de treballar en el tractament de les dades (o deixar molt de banda) i dedicar-se plenament a realitzar tasques que entren dins de la seva assignació. A més a més, si el sistema d'emmagatzemament ofereix algun tipus de servei de control d'error (per exemple el control sobre la inserció de valors negatius en les vendes per unitat d'un producte) podria arribar a reduir considerablement la feina de revisió que hi ha darrere de cada anàlisi deguda a la poca confiança que generen les dades enviades per part dels diferents clients d'IKI Media Communications S.L. Un altre punt a tenir en compte és que, el fet de poder emmagatzemar un històric complet i accessible (fet que ara resulta impossible a base d'excels) incrementarà el nombre de dades a usar en les anàlisis i per tant, resultarà molt més senzill d'aconseguir augmentar l'encert d'aquestes.

Què vol? Volen un sistema que permeti l'obtenció de dades tractades directament i filtrades i agrupades segons les necessitats momentànies que es tinguin en cada moment i en cada anàlisi. Amb aquest fet pretenen reduir la seva càrrega i tenir més temps per a l'estudi posterior dels resultats. També pretenen poder fer anàlisis amb més factors i més tipus d'anàlisis que permetin donar conclusions més exactes i prediccions més encertades. Un altra petició recau directament sobre el tractament de dades que volen deixar de fer i assegurar que funciona per si sol. A més a més, tenen intenció de millorar el seu servei mitjançant el dashboarding, fet que en cas de voler-se mantenir actualitzat, necessita que prèviament s'hagi implementat un sistema de bases de dades que doni accés a les dades en temps real (1 dia) i no s'hagi d'actualitzar manualment.

Com comunicar-nos? Comunicació directa, cara a cara diària per facilitar informació sobre l'avenç del projecte. Dubtes, consells i comunicació via mailing (empresa) per deixar constància de decisions crítiques del sistema en les que estiguin involucrats.

Estat actual? Els estadístics actualment es troben sobrecarregats, tenen més feina de la que poden arribar a fer en el temps de treball ja que les seves tasques assignades no són les úniques tasques que acaben realitzant. A més a més, a causa d'aquest fet, tot el departament de Data Science es troba saturat i no pot assumir més feina de la que té, fet que obliga que actualment existeixi la necessitat de créixer en personal per assumir més càrrega. Existeix també una por real entre tots els components del grup d'estadístics a les bases de dades per males experiències en casos anteriors basades en la poca flexibilitat.

11.5 Conclusions stakeholders

Concloem doncs amb l'anàlisi exhaustiva dels principals stakeholders. Visualitzem que un fet que juga en el nostre favor és que, en el cas dels stakeholders més rellevants del nostre projecte, tots obtenen benefici de la realització d'aquest projecte, fet que, acaba ajudant al fet que molt probablement aquest projecte es trobi amb menys problemes que no pas si existissin casos en què algun dels stakeholders afrontés una pèrdua a causa del projecte. A més a més, per la necessitat real de millorar la situació actual que veiem reflectida, podem gairebé assegurar que existirà col·laboració en el projecte per part de tots.

Observem també que la comunicació (necessària per a la col·laboració) amb la majoria dels stakeholders està assegurada i podrà ser feta cara a cara, fet que facilitarà molt que les consultes puguin ser més freqüents i que, generalment, no hi existeixi temps de demores entre respostes. Pel que fa als stakeholders amb els que no es pot arribar a realitzar comunicació directa cara a cara de manera senzilla i freqüent cal remarcar que, són stakeholders que realment en el scope del projecte no haurien de ser contactats, per tant, no resulta un problema representatiu pel projecte.

Pel que fa a l'estat actual de tots els stakeholders tornem a trobar indicis determinants de la necessitat de canvi tot i que, en cap cas, existeix un malestar directament expressat per cap dels casos estudiats, en tots els casos apareixen clars senyals de la necessitat que tenen cada un d'ells de millorar la seva situació. Bé sigui per qüestions purament de negoci (client, equip directiu) com per qüestions de millora de la situació laboral i professional dins de l'empresa (estadístics i analista programador). A més a més, descobrim també un interès fins ara no enunciat que consisteix a fer que el sistema dissenyat funcioni com a base del desenvolupament de cara a futures necessitats com ara eines, dashboards i aplicacions.

Un altre fet del qual ens n'adonem és que, en gairebé totes les descripcions de stakeholders realitzades apareixen les paraules: càrrega, error, tasques i freqüència. De cara a una possible solució, això, ens deixa entreveure que, el principal benefici que haurem d'aconseguir per a satisfer als diferents stakeholders rellevants serà oferir una millora relacionada amb aquest seguit de paraules. És a dir, s'haurà d'aconseguir millorar la situació en aquests àmbits aconseguit que es redueixi la càrrega de treball suposada pel tractament de les dades i l'emmagatzemant d'aquestes, que es disminueixi l'error a causa del factor humà (fet que alhora hauria de reduir la càrrega tant econòmica com de treball), que es realitzi una definició clara de les tasques a realitzar per cada treballador i que existeixi un procés estandarditzat que permeti una gestió més àgil de les anàlisis i finalment, que degut a la reducció de temps en la realització de tots els processos es pugui oferir una major freqüència d'entrega del servei oferit.

Per tant, tot això ens dóna una referència general de les necessitats del projecte i una referència específica de què necessita cada stakeholder rellevant que ens permetrà determinar quins hauran de ser els requisits mínims perquè totes les parts estiguin d'acord amb la proposta. Tot i això, realitzant aquesta anàlisi exhaustiva ens n'adonem també de què el tipus de stakeholder dels que parlem són de perfils molt diferents i que, per tant, gran part d'aquesta informació recopilada no podrà ser usada per tots els casos, sinó que s'haurà de discriminar en quins punts del projecte resulta representativa i important l'opinió de cert stakeholder i en quina no. Per exemple, en el cas de l'equip directiu, l'efecte que haurà de tenir sobre les decisions envers la usabilitat o la tecnologia usada pel projecte no haurà de ser rellevant a no ser que aquestes decisions acabin suposant un increment en el cost final del desenvolupament o el manteniment.

Així doncs, i per determinar més concretament l'àmbit en el qual s'ha de tenir en compte cada un dels stakeholders, s'ha decidit usar un mètode de classificació basat en la teoria presentada per 'Corporate Excellence'. Aquest mètode, novament es basa en la graficació dels diferents stakeholders per tal de situar-los en un espai que els defineixi.

En aquest cas la gràfica estarà formada per tres conjunts que intersecciones representats per cercles i vinculats a un atribut que pretén definir la posició de l'stakeholder. Els tres atributs que defineixen els conjunts són:

- *Poder:* pot influir en altres per prendre decisions que no perdrien per compte propi.
- *Urgència:* la relació amb l'interessat està marcada pel temps i és clau per a l'empresa.
- *Legitimitat:* té la capacitat moral o legal d'influir sobre el comportament de l'empresa.

Procedim doncs a col·locar cada un dels stakeholders en el conjunt que creiem que pertany per tal de posteriorment ser capaços de veure quin tipus de stakeholder representa i per poder valorar en quins àmbits del treball serà potencialment necessari tenir-lo en compte i en quins seran més prescindible.

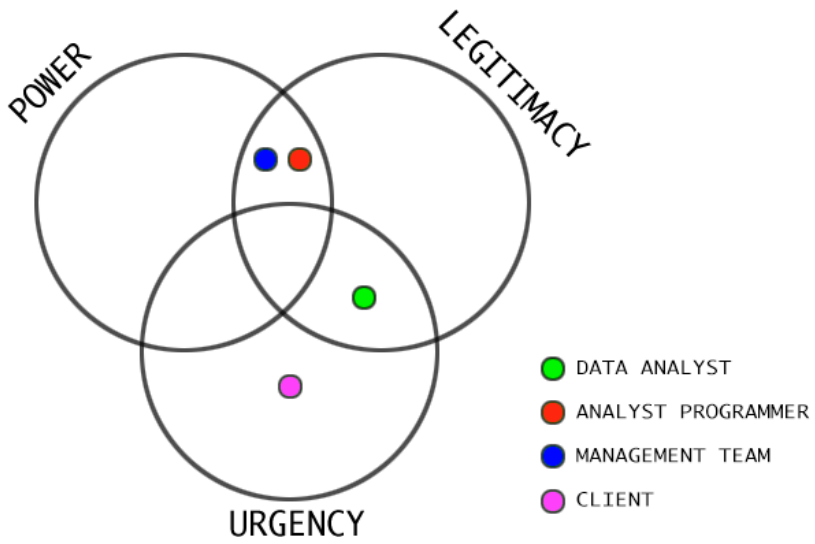


Figura 11.2: Esquema de Corporate Excellence

Veiem doncs en la figura 11.2 que hi ha 2 stakeholders amb poder, 3 amb legitimitat i 1 amb urgència. La justificació per al posicionament de cada un d'ells és la següent:

- *Equip directiu:* Està posat a poder, perquè pot prendre decisions que alterin el comportament dels altres i a legitimitat perquè té la capacitat moral d'influir en l'empresa, ja que, pot per exemple, triar l'ordre de rellevància dels diferents projectes a realitzar.
- *Analista programador:* Es troba a poder, perquè pot prendre decisions que facilitin l'adquisició d'un software o un altre fent que s'alteri el funcionament executiu de l'empresa i a legitimitat perquè té la capacitat d'influir en l'empresa, ja que, és el que porta tot el sistema de funcionament informàtic i per tant, qualsevol decisió seva, en ser la persona qualificada en aquest àmbit, serà portada a terme per tot l'equip.

- *Estadístic:* S'ha col·locat en urgència, ja que, la seva feina és en projectes temporals i és clau per a la captació de clients de l'empresa. També es troba a legitimitat, ja que, al ser un perfil tècnic, les seves decisions quant a inversió i necessitats en el Data Science solen ser tractades com a vàlides i considerades per l'equip directiu.
- *Client:* El client tan sols s'ha posat dins d'urgència, ja que, la seva implicació és temporal (una campanya) i clau al representar la font d'ingressos de l'empresa.

Un cop justificat el posicionament, aprofitem la descripció donada pel mateix 'Corporate Excellence' per determinar en quins àmbits han de ser rellevants cada un d'ells, presentat en la figura X.

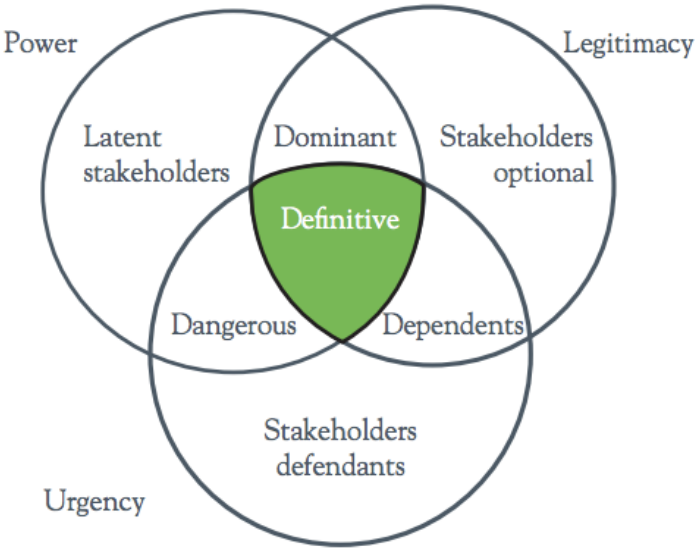


Figura 11.3: Definició de Corporate Excellence

Descobrim que tenim l'equip directiu i l'analista programador en una posició dominant i que per tant s'hauran de tractar especialment en aquells casos en els quals s'hagin de prendre decisions d'elecció de software, de tecnologies, etc. És a dir, en els casos en els quals pot suposar una inversió o en els casos que són determinants per al futur de l'empresa. En canvi, pel que fa a consultes per la operatibilitat del sistema, és millor consultar als estadístics, ja que, es troben en una situació de dependència. És a dir, que el software/sistema generat acabarà fent variar la seva forma de treball. Per tant, resulta coherent que per tots els temes més específics de l'ús del sistema, aquest stakeholder sigui més considerat, ja que, serà l'usuari final i per tant, el que haurà de sentir-se còmode amb el funcionament de la solució. Finalment tenim el cas del client que està situat en la posició de defensor del stakeholder i que actua com un stakeholder extern al qual no hauríem de tenir en compte de cara al desenvolupament.

12

Estudi de Context: Anàlisi de processos

Com s'ha explicat anteriorment, existeix una gran importància en la determinació dels processos que es segueixen pel tractament de les dades a usar i l'ús d'aquestes. Aquest fet és deriva de què actualment el model de processos és manifesta com a inexistent, tot i que inconscientment es porta un flux de treball amb nivell de rigorositat baix. Per tant un dels primers passos per entendre el domini, amb l'objectiu de poder començar a plantejar els requisits, serà entendre el mètode que permet realitzar la feina i en cas de trobar-nos amb indeterminacions o possibles millores plantejar-les i amb la col·laboració dels diferents stakeholders establir un diagrama de flux que expressi el funcionalment real actual.

Val a dir que una de les llibertats que han estat ofertes és que en cas de trobar un mètode que aporti millors resultats o faciliti els processos, consensuem amb les diferents parts afectades la integració de la nova metodologia i basem el disseny i el desenvolupament del projecte en la millora proposada. Aquest fet fa que, aquest anàlisi inicial tingui un pes més elevat dins del projecte, ja que serà un primer contacte amb la metodologia que aportarà una visió bàsica de possibles millores integrables.

Comencem doncs amb l'ànalisi de processos a realitzar. Per a fer-ho és realitzará un seguit de diagrames de flux mitjançant l'eina Draw.io proporcionada per Google i amb els estandards gènerics dels diagrames de flux.

Com a primer pas s'ha de determinar el 'scope' d'estudi del diagrama de flux que en aquest cas serà des de el moment en el qual les dades són recollides o rebudes fins al moment en què es finalitza l'estudi de les mateixes i es presenten els resultats. Aquest scope resulta ser més gran del que realment tracta aquest projecte (que es básicament la recollida, tractament i emmagatzament de les dades) però és óptimm, ja que, ens permetrà determinar quins seran els casos d'ús de les dades emmagatzemades i per tant, donar una solució que s'adapti més a les necessitats. Cal tenir en compte que aquest procés no acaba amb el valor d'aquestes dades, que idíl·licament s'haurien de conservar per a futurs estudis, però l'estudi no inclourà aquest cas pel simple fet que es dóna per suposat (basat en contacte directe amb el personal que usa les dades) que un cop realitzat el primer tractament de les dades, són usables i creuables directament en futurs escenaris és a dir, que és guarden com a històric.

Un cop determinat el 'scope' passem a presentar un diagrama inicial que ens permeti identificar les principals subprocessos a treballar. Aquest no era l'objectiu inicial per a la realització d'aquesta tasca però degut a problemes en l'acord de reunions (resultaven ser més reunions de les que l'equip podia assolir), s'ha decidit modificar el metode de realització passant a primerament realitzar un esquema de processos genèric que és comentarà amb tots els stakeholders alhora fins que a base d'iterar sobre el mateix, s'arribi a un diagrama que consideri les descripcions de tots. L'objectiu d'aquesta primera representació no és en cap cas ser plenament fidel a la realitat sinó aproximar-nos al domini per a una millor comprensió i facilitar que és puguin detectar les diferents subtasques amb vista a en un futur poder definir-les.

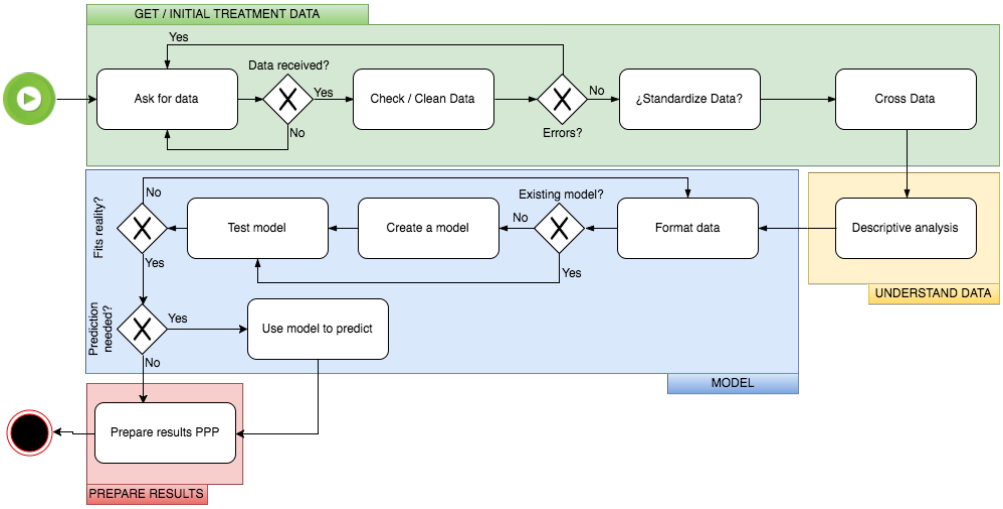


Figura 12.1: Diagrama de flux inicial

Aquest diagrama ha estat fruit d'un primer contacte amb els diferents treballadors que realitzen aquest procés i d'una segona reunió que ha permés verificar que el diagrama que s'ha generat encaixa amb les diferents visions del procés que és porta a terme. Tenim doncs un primer diagrama de flux genèric i passem a definir els diferents subprocessos detectats fins al moment per especificar-lo més concretament:

Ask for data: Les dades a tractar poden provenir de les 4 fonts diferents que s'han explicat anteriorment (IOPE, INV, KM, ESP). En el cas de les dades del tipus ESP, la tasca de realitzar les peticions de dades és duta a terme per un nombre elevat de persones, ja que les dades usades en els estudis són dades que sovint poden haver estat generades o demanades en altres departaments. Pel que fa als altres casos, és a dir a IOPE, INV i KM, observem que són recollides directament un sol cop.

Check/Clean data: En aquest punt es dóna per descomptat que ja s'han recollit les dades. La tasca corresponent ara és la de netejar les dades. Això representa genèricament tractar dos tipus d'errors freqüents:

- *Errors en les dades:* Aquests errors es donen en casos en què després d'una primera inspecció de les dades, es troben valors incomprensibles o que han estat mal introduïts. Per a solucionar-ho es parla amb el responsable de les fonts de dades per tal de corregir-ho o entendre les desviacions.
- *Missing values:* En aquest cas, els errors són deguts a incompleteses en les dades que s'han rebut. En aquest cas també es parla amb el responsable de les fonts, però aquest cop, en cas de no poder tornar a rebre-les corregides s'ha d'usar metodologies de 'filling missing value' per tal de completar-les i mantenir la base de dades completa.

¿Standardize data?: Un cop s'arriba a aquesta tasca, es dóna per descomptat què tenim unes dades sense errors i completes. La feina a realitzar en aquest punt és agafar les diferents dades recollides de diferents fonts i estandarditzar-les per tal d'aconseguir un mateix format de per a cada valor i mantenir un estàndard entre les dades. Per exemple, podem trobar expressada de diferent manera la columna "data"depenent de la font, l'objectiu doncs serà trobar un format d'aquest paràmetre que permeti estandarditzar-ho i tenir dades coherents i cohesionades entre les diferents fonts. Cal notar que aquesta

tasca es troba entre interrogants, aquest fet és degut al fet que a visió personal, aquests estàndards són poc definits (no s'ha trobat documentació al respecte) o arbitraris. Per tant des de el meu punt de vista aquest punt és pràcticament inexistent i és barreja plenament amb el següent que s'explicarà.

Cross Data: Una vegada acabada l'etapa de tractament de les dades per separat, es passa al punt en el qual s'ajunta les dades recollides per formar un arxiu contenidor de tota la informació centralitzada. Aquest procés es realitza de manera rudimentària mitjançant Excel i encreuaments semi manuals basats en macros i fórmules programades del mateix software. Sovint aquesta tasca requereix ser totalment manual, ja que, com que no existeix un estàndard entre les diferents fonts de dades (com en el punt anterior s'ha comentat), els scripts existents no donen bons resultats a l'encreuar les dades correctament. En aquest pas acabem aconseguint un fitxer amb totes les dades centralitzades i correctament encreuades.

Descriptive Analysis: Tenint ara les dades encreuades i sense errors visibles, es passa a un procés en el qual es realitzen anàlisis descriptius que permetran entendre les dades a treballar. Per a aquest procés les dades es passen a format .csv per tal de facilitar-ne la lectura al software R. Un cop passades les dades a R es realitzen un conjunt de gràfiques com histogrames, boxplots, barplots, gràfics de lineas... Quan es té el conjunt de representacions gràfiques i informes numèrics suficients (i.e. estudis de correlació chi-squared), es passa a una fase en la qual s'estudien per tal de comprendre les dades a les quals l'estudi s'haurà d'enfrontar. Sovint en aquesta fase es requereixen interaccions personals per tal de poder transmetre coneixement dels diferents departaments, ja que sovint algunes accions publicitaries realitzades afecten les dades pel que resulta de vital importància la comunicació d'aquest tipus d'accions per a facilitar la comprensió de les dades a treballar.

Format Data: Un cop s'han entès les dades que s'hauran de treballar es passa a preparar les dades de l'execució del model. En aquest pas, el que es realitza són quatre accions bàsicament:

- *Donar el tipus adequat a les dades:* En aquest pas en verifica que el tipus de les dades és el correcte i en casos específics es preparen els labels que actuaran de resultat.
- *Tractament premodelització:* Sovint les dades requereixen tractament abans de passar-les pels models per tal d'aconseguir un millor ajust (i.e. dividir les numèriques per la mitjana per reduir el rang de variabilitat, realitzar acumulacions de variables sobre una nova columna, etc.) o simplement perquè alguns models no accepten tipus específics i s'han de modificar les dades per a poder-les entrar (i.e. categòriques passades a binaries, numèriques passades a rangs, etc.).
- *Afegir columnes extra:* En alguns casos, el model requerirà l'afegiment d'informació extra per tal de donar més ajust al model i serà en aquest moment en el qual s'afegiran manualment aquestes dades.
- *Selecció de paràmetres:* Finalment en aquest procés gairebé tots els casos es selecciona un subconjunt de les dades totals per a aplicar sobre el model.

Create Model: Aquest subprocés tan sols es porta terme quan un tipus de modelatge no ha estat prèviament implementat i es requereix implementació d'aquest. En cas d'haver-se de realitzar el que es fa és generar un codi a mode de llibreria/funció (depenent de la importància del model) de R que permeti executar el tipus de model per a qualsevol dada d'entrada. És a dir, es busca que el resultat sigui un script reusable per a tots els casos d'execució d'aquell model per a futurs anàlisis.

Test Model: Suposem doncs en aquest punt què es té un conjunt de dades sense error, estandarditzades i amb un format correcte per a l'execució del model i una implementació d'una llibreria/funció de R. En aquest punt doncs la idea és executar el model amb les dades i estudiar-lo per veure si els resultats obtinguts poden considerar-se correctes o no encaixen amb la realitat i per tant s'ha de buscar un altre model o subconjunt de dades per a disminuir l'error. Un punt a tenir en compte és que sovint aquest procés requereix d'un anàlisi basat en paràmetres diferents de l'error per a la comprensió dels resultats dels models.

Prepare ppt results: Finalment s'executa aquesta acció. En aquesta acció s'agafen tots els resultats obtinguts durant els subprocessos anteriors (anàlisi descriptius, models, prediccions, etc.) i es prepara una presentació per a poder explicar al client els resultats de l'estudi realitzat. Aquestes presentacions segueixen un patró molt similar i tenen un estil definit per tal de mantenir coherència d'imatge de l'empresa. Un cop acabats de preparar, s'afegeixen comentaris explicatius de tots els resultats coma revisió final i per a afegir valor al producte final de l'anàlisi, és a dir, es pretén que el client no tan sols tingui la informació explicada el dia de la reunió sinó que mantingui un petit resum explicatiu escrit als resultats que s'entreguen.

Aquestes definicions han estat comentades amb l'equip per tal de veure si eren adequades a la realitat i les han acceptat en gran part afegint petites consideracions de casos exepcionals que s'ha decidit no tractat per part meva ja que, sota el meu punt de vista, no són casos que realment necesitessin ser tractats com una exepcionalitat sinó que el fet de no tenir un flux determinat i canviant ha tolerat que casos que eren incluibles dins d'un flux estandaritzat fossin tractats com casos especials.

Seguim, un cop aconseguida aquesta definició inicial vàlida dels subprocessos, realitzant un estudi en profunditat de cada subprocés per a l'obtenció de les metodologies usades en cada un d'ells i també per a obtenir una representació real dels arxius implicats del procés. Recordem que el funcionament actual està basat en l'actualització d'arxius .xls i .csv. El resultat del següent pas de l'estudi per tant ens permetrà aconseguir una definició concreta del procés i s'espera que posteriorment mitjançant una fusió entre les definicions dels subprocessos s'aconsegueixi el diagrama de flux complet.

Destaquem abans de començar què es farà servir un sistema de colors per identificar els arxius que apareguin en els diagrames de flux. S'usarà el color vermell per a identificar els arxius / bases de dades de input que no són controlats per l'empresa, el color taronja per determinar els arxius intermedis de cada subprocés que acaben eliminats o que no són directament treballats per altres subprocessos i verd per a determinar els outputs de cada subprocés.

12.1 Ask for data

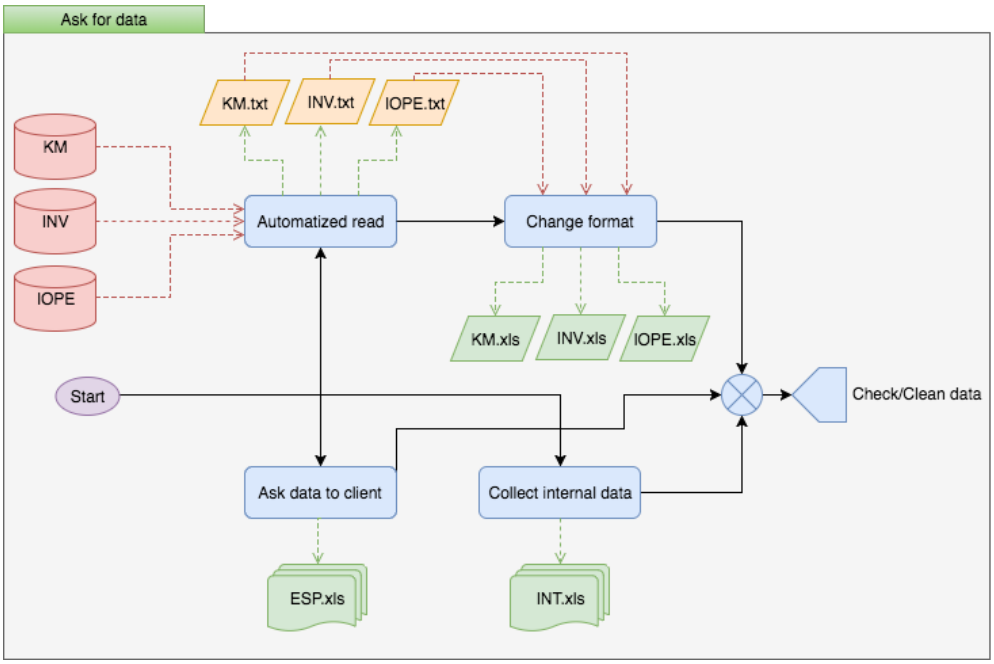


Figura 12.2: Diagrama de flux: Demanar dades

Aquest doncs és el diagrama (figura X) que ens permetrà definir el procés de demanar i recollir les dades. És observable en el diagrama que existeixen 3 fonts de dades estàtiques que han estat explicades amb anterioritat, aquestes són KM, INV i IOPE. Aquestes fonts de dades són llegides automàticament i les consultes a les mateixes són emmagatzemades planament en .txts. Posteriorment se'ls canvia el format i se'ls dona un format llegible mitjançant Excel i macros ja preparades. Les altres fonts de dades existents que s'han de recollir no són tan trivials.

Comencem pel cas de les dades ESP, aquestes són les dades que el client proporciona de la seva facturació, impacte, vendes, unitats, etc. Al ser dades tan independents, la comunicació per a la recollida sol ser forçament amb diferents departaments dels clients i les dades per tant solen ser lliurades en més d'un .xls. I acabem parlant de les dades internes a recollir. En aquest cas trobem quelcom similar al cas anterior, les dades internes són gestionades per diferents departaments i per tant rebem un conjunt de .xls que contenen les dades separades per àmbit de treball de l'empresa.

Concloem doncs que com a resultat d'aquest subprocés s'obtenen tres .xls amb les dades respectives de les fonts estàtiques i una quantitat n de .xls provinent de les dades de client i de les dades internes depenent de l'estudi a realitzar.

12.2 Check/Clean data

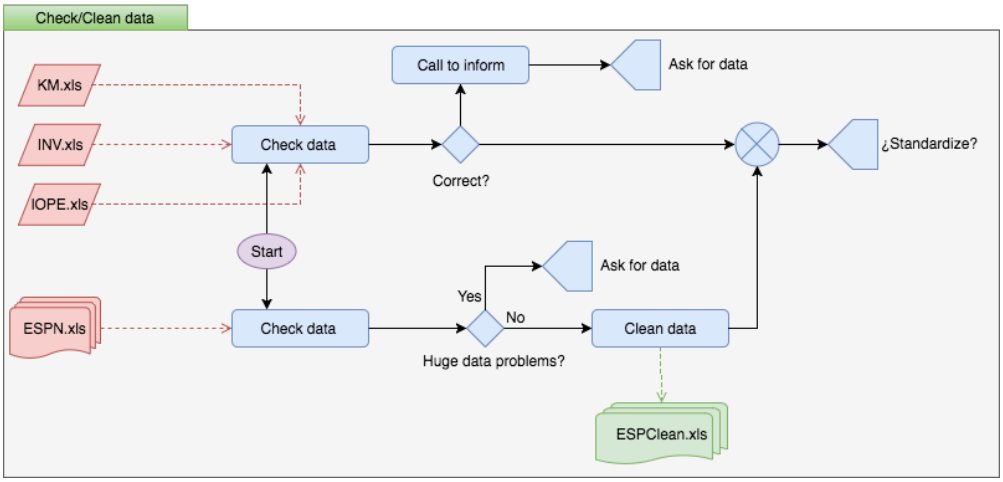


Figura 12.3: Diagrama de flux: Revisar/Netejar dades

Un cop recollides les dades, s'ha de comprovar que aquestes són correctes. En el cas d'estudi que estem tractant s'han de fer dues comprovacions paral·leles, ja que existeixen dues classes de dades: les provinents de serveis confiables (KM, INV i IOPE) i les provinents de fonts de dades no fiables (ESPN i INTN). El primer pas a portar a terme amb totes les dades és una comprovació extensiva per a detectar possibles errors de la font, com ara: manca de dades en cert període, valors erronis, etc.

Comencem parlant del cas del grup de dades fiables. Si es troben errors, es notifica la font de dades perquè ho solucioni i es tornen a demanar les dades. En cas contrari s'espera a tenir les dades ESPN i INTN tractades i revisades per passar al següent subprocés.

En canvi en el cas de les dades internes o les específiques en el cas de trobar molts errors simplement es demanen de nou unes dades correctes. Seguidament un cop les dades rebudes són suficientment correctes es passa a realitzar una neteja de les mateixes per a solucionar problemes menors que puguem trobar en aquestes. Finalment s'espera a tenir les dades KM, INV i IOPE tractades i revisades per a passar al següent subprocés.

12.3 ¿Standardize?

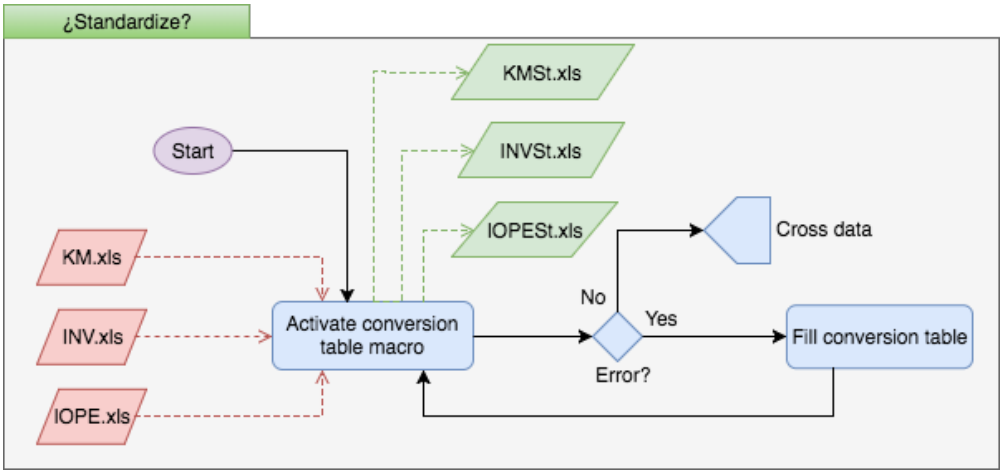


Figura 12.4: Diagrama de flux: Estandarditzar dades

Passem ara a un procés que com anteriorment ha sigut mencionat, no sembla del tot definit i que per tant, s'ha basat en les descripcions donades del mateix perquè no existeix documentació relativa al tema ni als estàndard a aplicar. En aquesta tasca per tant es pretén aconseguir un estàndard per a totes les fonts de dades confiables. Per a fer-ho s'executa una macro sobre els diferents arxius .xls que modifica els valors existents per les conversions existents en una taula que conserva una relació entre els valors reals i la seva conversió a estàndard.

Un cop executada aquesta macro, es comprova que no hi hagin aparegut nous valors a la taula de conversió automàticament (la taula de conversió s'actualitza automàticament en cas de trobar valors que no sap convertir i els deixa registrats). En cas d'haver-hi errors, s'omple la taula i es torna a executar la macro.

Com a resultat d'aquest procés obtenim els arxius KMSt, INVSt i IOPESt. Que representen ser la conversió a estàndard dels KM, INV i IOPE.

12.4 Cross data

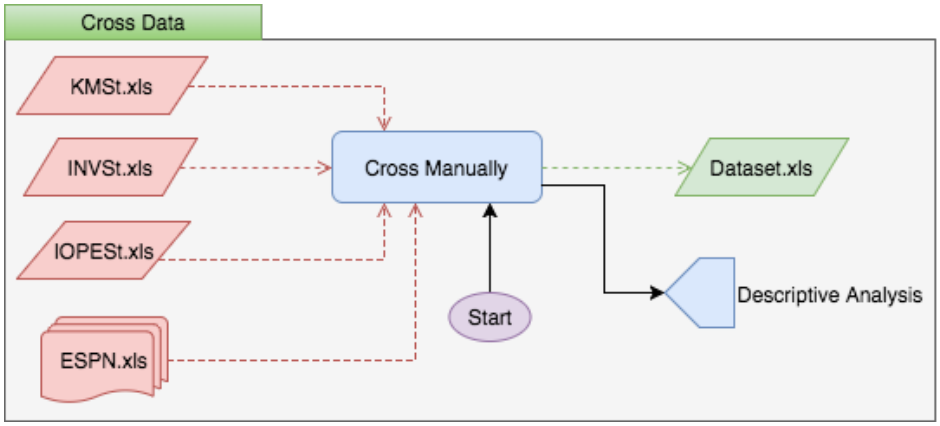


Figura 12.5: Diagrama de flux estandaritzar dades

Un cop estandarditzades les dades es realitza un encreuament. Entendrem ara perquè segons el nostre punt de vista no existeix una estandardització real de les dades. El què es realitza en aquest procés és encreuerar manualment les dades per tal d'aconseguir un sol document amb totes les dades necessàries. D'aquí sorgeixen certes preguntes que ens faran acabar descobrint que quelcom no encaixa:

- *Si tenim valors estandarditzats, per què no usar macros per creuar les dades?* Aquesta pregunta però podria tenir resposta, el fet és que no existeix un estàndard per les dades específiques sinó tan sols per les fonts fiables.

- *I doncs perquè no creuar les dades fiables amb macros i tan sols ajuntar les altres dades manualment?* És aquesta la pregunta a la que no trobem una resposta clara. I el fet pel qual justifiquem que el subprocés d'estandardització no existeix i que és en aquest procés en el qual es canvien valors i es creuen arbitràriament les dades depenent de la informació rebudes i l'objectiu.

Per tant, el què es realitza en aquest procés és a partir dels arxius llegits, formar-ne un manualment fent encaixar les dades manualment i realitzant aquest tractament d'acord ambl cas d'estudi.

12.5 Descriptive analysis

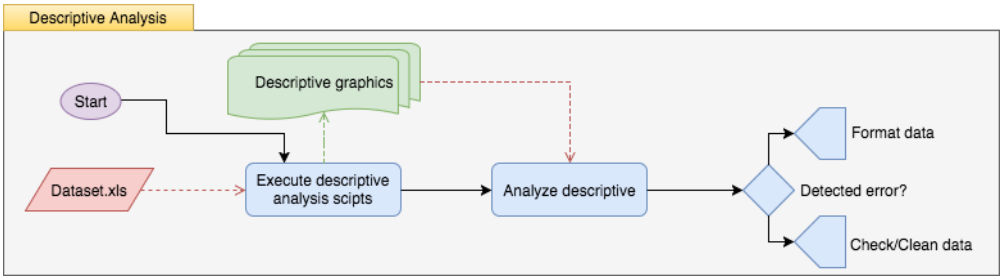


Figura 12.6: Diagrama de flux: Anàlisi descriptiu

Passem ara a un nou paquet de subprocessos, en el que ja donem per suposat el final del tractament inicial de les dades (també anomenat preprocessing). En aquest moment el que s'ha de tractar és d'entendre les dades amb les quals posteriorment s'haurà de treballar. Per tal de fer-ho es tenen uns scripts de R predeterminats que ens permetran autoexecutar-se i generar els anàlisis descriptius bàsics (boxplots, histograms, barplots, etc.) entre altres gràfiques per a facilitar la comprensió de dades a treballar.

Un cop executat i generats tots els documents descriptius, es revisen per veure si es poden trobar errors que prèviament no fossin observables. Alhora en aquest procés s'intenta d'entendre totes les dades existents, és a dir, intentar de comprendre el perquè dels comportaments de les dades, de l'estructura d'aquestes, les seves distribucions, etc.

En el cas de detectar errors, es torna a comprovar i netejar les dades i en cas de no poder solucionar el problema, es tornarà a demanar dades per evitar els errors. En canvi si les dades són correctes, es canvia de paquet i es comença l'estudi de modelatge. Els resultats d'aquest subprocés seran tots l'anàlisi descriptius executats.

12.6 Format data

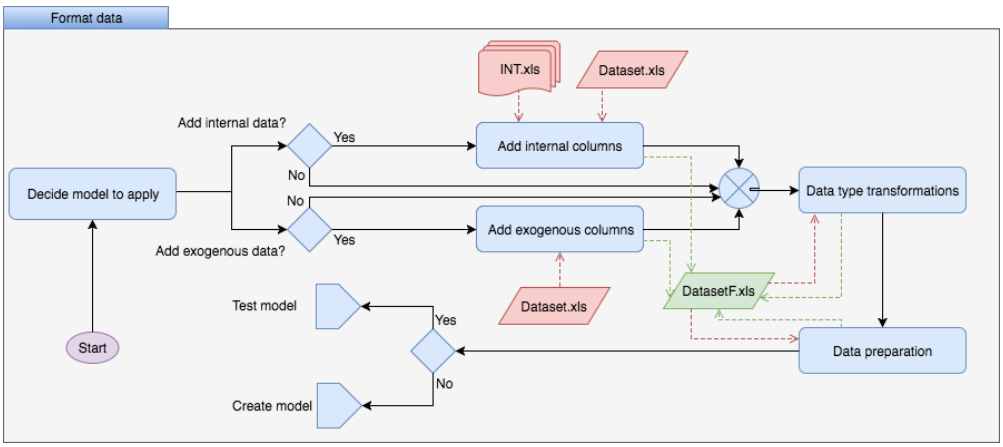


Figura 12.7: Diagrama de flux: Preparar dades pel model

Entrem ara doncs al paquet d'estudi basat en models. Cal definir que aquest paquet resulta ser iteratiu, és a dir, serà un paquet sobre el qual molt probablement repetirem execucions del seu flux. Aquest s'inicia amb l'apartat en què és pretén de donar format a les dades per a la futura execució. En aquesta tasca doncs l'objectiu és aconseguir un dataset amb la informació que volem que s'apliqui al model.

Primerament el que s'ha de fer és decidir sobre quin tipus de model usarem les dades, resulta ser imprescindible pel fet que molts models actuen amb millor precisió en dependència de les dades introduïdes. Un elegit el model s'ha de decidir si afegir informació exògena o interna al dataset per a tenir més paràmetres per l'estudi, en cas afirmatiu, s'han d'afegir les columnes desitjades.

Posteriorment, es realitzen canvis sobre els formats de les dades per tal de poder-les usar en cada model. Un exemple d'aquest procés és quan donada una dada categòrica s'ha de transformar en una explosió de columnes a una binaria per tal de poder ser usada per models que no accepten aquest tipus de dades. Un altre pas que es realitza és preparar les dades perquè donin un millor resultat, per exemple, dividint les variables numèriques per la mitja per tal de reduir el rang dels diferents paràmetres i millorar-ne els resultats.

Finalment, un cop tenim les dades preparades, passem a l'execució del model. En cas de ser existent, simplement és prova, en cas contrari, el script del model s'ha de crear.

12.7 Create model

Suposem doncs que el model encara no ha estat implementat. En aquest cas simplement s'ha d'implementar un codi tal que permeti executar el model amb qualsevol dataset. Recordem que la realització d'aquests scripts resulta ser complexa i que requereix un estudi previ de les dades que podran rebre, etc. No s'ha graficat el subprocés, ja que bàsicament és una sola acció de creació en la qual es crea el script, i per tant bàsicament ho realitza una persona tal com desitja i usant els recursos que trobi pertinents, com ara, estudi de les dades a usar, consultes web per a la cerca informació sobre els models, consultes de white papers, consultes de documentació referent a R, etc.

12.8 Test model

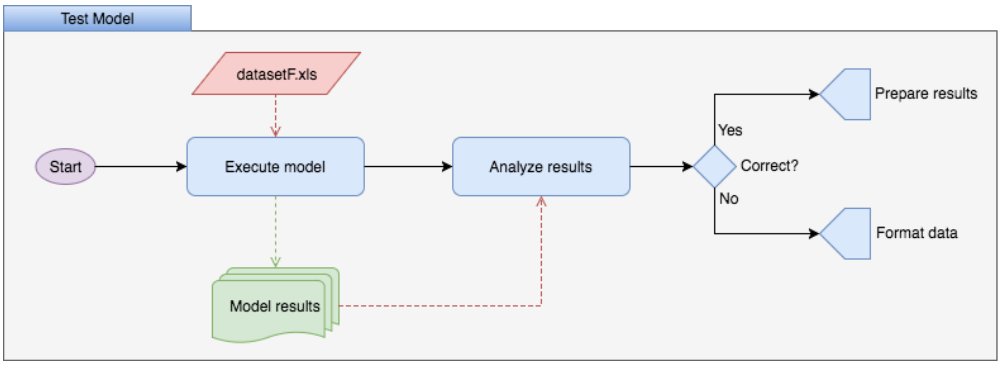


Figura 12.8: Diagrama de flux: Probar model

Finalitzant ara amb el package d'estudi del model, un cop es té un model executable i les dades per a executar-lo es passa al procés de prova del model. Per tal de fer-ho simplement s'executa el script de R amb les dades tractades anteriorment, és a dir amb el datasetF.csv.

D'aquest fet, se n'aconsegueixen els resultats del model executat. Un cop aconseguits el que s'ha de fer és analitzar els resultats per a decidir si resulten un estudi vàlid o introdueixen massa error no so són prou explicatius. Per tal de fer-ho sovint es requereix ser entès en el scope del estudi i haver consultat prèviament els anàlisis descriptius.

Un cop realitzat aquesta anàlisi si es decideix que és vàlid es passa a preparar una presentació de resultats i en cas contrari es torna a donar format a les dades per triar un altre model o provar amb un altre subset de variables generant així una iteració fins a trobar una solució.

12.9 Prepare ppt results

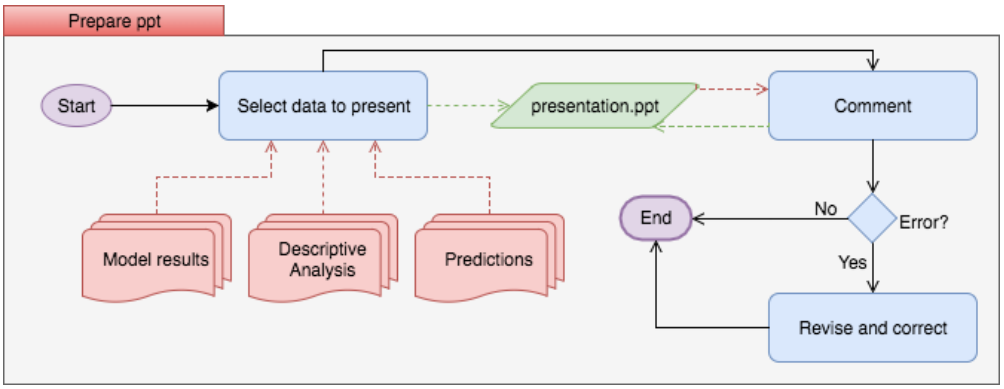


Figura 12.9: Diagrama de flux: Preparar presentació

Concloem el procés entrant directament a l'últim paquet de preparació dels resultats per a ser ensenyats. Per a tal motiu es realitza el subprocés de creació del powerpoint amb els resultats aconseguits. Per tal de fer-ho, primerament es realitza el powerpoint mitjançant tots els resultats reals aconseguits que són: anàlisi descriptius, models i prediccions en cas de ser existents. Un cop preparada la presentació, que serà feta segons un template que permet mantenir un format similar per a tota l'empresa, facilitant així el manteniment d'una mateixa imatge d'empresa, es passa a un altre actor diferent del que l'ha presentat perquè la revisi i afegeixi comentaris a cada slide que permetin afegir un valor a la presentació perquè el client pugui consultar-la tant en el moment de l'exposició com un cop acabada. En cas de detectar errors es solucionen bé sigui modificant i tornant uns quants subprocessos enrere (depenent de l'error) o actualitzant la presentació en cas de ser error de presentació. En cas contrari s'acaba el procés.

12.10 Diagrama processos final

Un cop realitzada una descripció extensa de cada tasca s'ha de concloure realitzant un esquema què encabeixi els diferents subprocessos prèviament exposats. Presentem per tant el següent diagrama de flux, que donem com a conclusió de l'anàlisi de les tasques prèviament exposat.

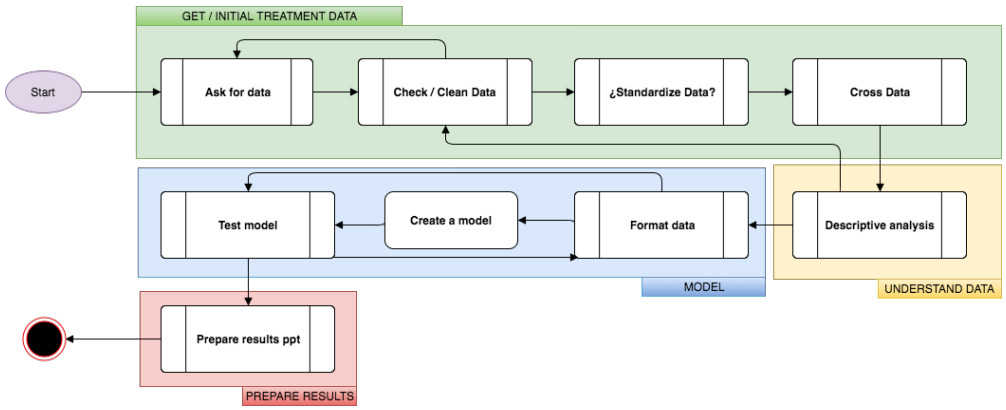


Figura 12.10: Diagrama de flux final

Veiem doncs que l'estudi realitzat és finalment l'encadenament prèviament plantejat en l'anàlisi en l'àmbit de subtasques realitzat. És important determinar l'existència dels 4 paquets, ja que determinen la fase en què ens trobem del desenvolupament i depenent del paquet els treballadors a realitzar la tasca haurien de resultar ser diferents.

13

Especificació de Requisits

Una vegada estudiat l'entorn en el qual ens trobem mitjançant l'estudi de context realitzat, és a dir, l'estudi de les fonts de dades, l'anàlisi del hardware i del software, l'estudi dels stakeholders rellevants en profunditat i la descripció del procés d'actuació, procedim a especificar els requisits i requeriments que el sistema haurà de complir per tal d'adaptar-se al context i millorar al màxim la situació actual. Cal tenir en compte que aquests requisits, per tant, estan basats en el bé comú de tots els stakeholders i contemplen el cas genèric per davant de l'específic és per això que poden existir objectius d'un stakeholder no complerts, ja que, de cara al bé comú no són rellevants.

L'especificació de requisits doncs ens permetrà determinar i especificar les necessitats i condicions de la solució a generar. Fent així que, un cop acceptats per part de totes les parts, quedi l'objectiu final del projecte definit i les condicions que haurà de complir per a ser un disseny vàlid.

Comencem doncs presentat els requisits funcionals i no funcionals que el sistema haurà de complir. Per fer-ho s'usarà la taula plantejada a l'estàndard IEEE 83 - 1998 i l'explicació del mateix realitzada pel 'Software Engineering Standards Committee'. S'ha prescindit de realitzar els primers punts descriptius que requereix aquest estàndard, ja que, en la seva majoria han estat realitzats extensivament en punts anteriors de l'estudi de context. Comencem doncs des de la pròpia definició de requisits, saltant-nos així, els punts d'introducció i descripció general a causa de la seva prèvia exposició. Per tal d'assegurar una descripció completa de cada requisit s'ha trobat una estructura en forma de taula que permet descriure'ls de manera estandarditzada (seguint la idea de Volere, però aplicant-hi les bases del IEEE 380) i estructurada (per a la realització d'aquesta taula s'han tingut en compte les bones formes proposades a l'IEEE 380). Les taules expressant els diferents requisits tenen els següents continguts, explicats a continuació:

- *Identificador:* Identificador del requisit, usat per a referenciar-lo en cas de necessitar-ho.
- *Nom:* Nom descriptiu del requisit. Haurà de permetre donar la idea bàsica que busca el requisit.
- *Característiques:* Exposició del requisit.
- *Descripció:* Descripció acurada del requisit exposat. Ha de quedar clar, el que queda dins d'aquest requisit i el que queda fora, és a dir, ha de definir el scope del requisit.
- *Stakeholder:* Stakeholders als que va dedicat aquest requisit.
- *Requisits No Funcionals (RNFs):* Requisits no funcionals vinculats a la realització d'aquest requisit.
- *Prioritat:* Prioritat d'aquest requisit. Aquest fet ens permetrà, en cas d'haver de discriminar un requisit per qualsevol problema, minimitzar els danys al resultat final.

A més a més, l'estructura del document serà basada en la trobada en els estàndards IEEE 830:
Procedim doncs a la realització de l'especificació dels requisits seguint les pautes prèviament explicades.

TABLA II

3.- Requisitos específicos
- 3.1 Requisitos funcionales
  - 3.1.1 Requisito funcional 1
    - 3.1.1.1 Introducción
    - 3.1.1.2 Entradas
    - 3.1.1.3 Procesamiento
    - 3.1.1.4 Salidas
  - 3.1.2 Requisito funcional 2
  - … … … … … …
  - 3.1.N Requisito funcional N
- 3.2 Requisitos de interfaz externa
  - 3.2.1 Interfaces de usuario
  - 3.2.2 Interfaces de hardware
  - 3.2.3 Interfaces de software
  - 3.2.4 Interfaces de comunicaciones
- 3.3 Requisitos de ejecución
- 3.4 Restricciones de diseño
  - 3.4.1 Acatamiento de estándares
  - 3.4.2 Limitaciones hardware
  - … … … … … …
- 3.5 Atributos de calidad
  - 3.5.1 Seguridad
  - 3.5.2 Mantenimiento
  - … … … … … …
- 3.6 Otros requisitos
  - 3.6.1 Base de datos
  - 3.6.2 Operaciones
  - 3.6.3 Adaptación de situación
  - … … … … … …

Figura 13.1: Estructura de descripció de requisits IEEE-830.

13.1 Requisits funcionals: Interfície externa

Aquest punt hauria de ser una descripció detallada de totes les entrades i sortides del sistema a dissenyar.

Identificador	RNF-IE01
Nom	El sistema ha de permetre tractar entrades de dades de la font de dades KM
Característiques	El sistema ha de ser capaç de realitzar l'inserció de les dades provinents de la font de dades KM.
Descripció	El sistema ha d'oferir la possibilitat d'inserció de les dades de KM tant de forma automàtica, com via API, com via arxius de text.
Stakeholder	Analista programador, Estadístics.
Prioritat	Mitjana

Taula 13.1: RF - IE01: Permetre tractar entrades de dades de la font KM

Aquest punt hauria de ser una descripció detallada de totes les entrades i sortides del sistema a dissenyar.

Identificador	RNF-IE02
Nom	El sistema ha de permetre tractar entrades de dades de la font de dades IOPE i INVE
Característiques	El sistema ha de ser capaç de realitzar l'inserció de les dades provinents de la font de dades IOPE i INVE.
Descripció	El sistema ha d'oferir la possibilitat d'inserció de les dades de KM tant de forma automàtica via arxius de text.
Stakeholder	Analista programador, Estadístics.
Prioritat	Alta

Taula 13.2: RF - IE02: Permetre tractar entrades de dades de la font IOPE i INVE

Identificador	RNF-IE03
Nom	El sistema ha de permetre realitzar entrades de dades de les fonts de dades específiques (client)
Característiques	El sistema ha de ser capaç d'integrar-se amb fonts de dades provinents de tipus específic.
Descripció	El sistema ha d'oferir la possibilitat d'integració amb dades específiques de cada client gràcies a mantenir una estructura d'atribut d'encreuament estandarditzada. Aquest fet no suposa que hagi de donar un sistema d'inserció sinó que s'ha d'oferir un sistema d'estandardització de les altres fonts que permeti a l'Analista programador generar noves estructures de dades que siguin fàcilment creuables amb les dades ja existents.
Stakeholder	Analista programador, Estadístics.
Prioritat	Alta

Taula 13.3: RF - IE03: Permetre realitzar entrades de dades de les fonts específiques (client)

Identificador	RNF-IE04
Nom	Obtenció datasets amb dades encreuades
Caracteristiques	El sistema ha de permetre realitzar l'extracció de datasets amb les diferents fonts de dades encreuades.
Descripció	El sistema ha de permetre aconseguir datasets amb totes les dades de les diferents fonts encreuades mitjançant el conjunt d'atributs d'encreuament triats i amb la granularitat desitjada (sempre tenint en compte les restriccions de les dades, d'unes dades setmanals no podem demanar una granularitat dia a dia). Aquests encreuaments hauran d'estar basats en els refactors que prèviament s'hauran d'haver determinat. Aquesta acció ídilicament en un futur haurà de ser feta mitjançant una eina amb interfície però per el scope d'aquest projecte (degut a què el que estem solucionant és el tractament i l'emmagatzament de les dades) haurà de ser amb queries SQL.
Stakeholder	Estadístics.
Prioritat	Alta

Taula 13.4: RF - IE04: Obtenció de datasets amb dades encreuades

13.2 Requisits funcionals: Funcionalitats

En aquest punt s'haurien d'exposar les funcionalitats esperades que porti a terme la proposta de solució. Són les següents.

Identificador	RF-F01
Nom	Detecció de necessitat de refactor
Característiques	El sistema ha d'informar a l'usuari, un cop realitzada una inserció de dades, de les possibles necessitats d'estandardització de dades prèviament no refactoritzades que apareixin.
Descripció	El sistema haurà d'avisar a l'usuari de qualsevol dada mitjançant les quals es pugui realitzar l'encreuament que hagi aparegut per primera vegada amb l'objectiu que l'usuari pugui mitjançant una interfície bàsica decidir quina haurà de ser la conversió d'aquell paràmetre per al encreuament. Per tant, la notificació haurà d'oferir un accés directe a la modificació de les conversions ja existents del sistema.
Stakeholder	Analista programador i estadístics.
Prioritat	Alta

Taula 13.5: RF- F01: Detecció de la necessitat de refactor

Identificador	RF-F02
Nom	Actualització dels refactors existents
Característiques	El sistema ha de permetre a l'usuari modificar les conversions a realitzar de cara als encreuaments
Descripció	El sistema ha de donar la possibilitat a l'usuari de modificar mitjançant una interfície gràfica les conversions dels atributs d'encreuament de les dades. Resulta important que l'històric de canvi d'aquesta conversió quedi registrat amb justificacions de les diferents preses de decisió. Les conversions a realitzar hauran de ser tractades client per client, ja que, en molts casos diferents clients entenen els mercats de maneres diferents. Els refactors s'hauran d'emmagatzemar i seran usats per a tots els encreuaments.
Stakeholder	Analista programador i estadístics.
Prioritat	Alta

Taula 13.6: RF - F02: Actualització dels refactors existents

Identificador	RF-F03
Nom	Consulta dels refactors existents i històric
Característiques	El sistema ha de permetre a l'usuari consultar els refactors existents i l'històric de cada un dels refactors
Descripció	El sistema ha de donar una interfície que permeti consultar els valors dels refactors, amb el màxim de filtres per a facilitar la cerca d'aquests. A més a més, resultarà necessari mostrar també, en cas de ser necessari l'històric de canvi de cada un dels refactors. Aquest requisit es justifica, ja que, durant la realització de les anàlisis resultarà important en cas de dubte poder, consultar les agrupacions realitzades.
Stakeholder	Estadístics.
Prioritat	Alta

Taula 13.7: RF - F03: Consulta dels refactors existents i l'històric

Identificador	RF-F04
Nom	El sistema ha de generar un log d'activitat dels usuaris
Característiques	El sistema ha generar un conjunt de dades que permeti detectar els canvis que cada usuari ha realitzat.
Descripció	El sistema ha de permetre identificar quines accions s'han realitzat i qui les ha realitzat. Amb això es pretén poder portar un control sobre el qual s'està fent en les dades per tal de, posteriorment i en cas de no entendre certs procediments realitzats, poder preguntar a l'usuari que ho ha portat a terme el perquè d'aquella decisió. Aquest requisit per tant també força al fet que hagi d'existir un sistema de comptes d'usuari que haurà de seguir la legislació vigent presentada en el BOE.
Stakeholder	Analista programador, Equip directiu.
Prioritat	Alta

Taula 13.8: RF - F04: Generar un log d'activitat dels usuaris

13.3 Requisits no funcionals: Performance

Aquesta subsecció ha d'especificar tant els requeriments numèrics estàtics com els dinàmics que es troben en el programari o en la interacció humana amb el programari en general.

Identificador	RNF-P01
Nom	Les tasques d'inserció del sistema no han d'aturar el treball de l'usuari amb el sistema.
Característiques	Les tasques d'inserció de dades i lectura de les mateixes no poden aturar el treball de l'usuari amb el sistema durant un temps perllongat
Descripció	El sistema ha de poder seguir-se usant durant la inserció. Resulta important que, tot i estar realitzant la lectura de les dades i el tractament d'aquestes, el sistema no s'aturi i l'usuari pugui seguir interaccionant amb ell. D'aquesta manera suposem que no pot existir una pausa superior als 2 segons a partir de l'inici de la inserció i que, en un temps no superior als 10 segons, les taules de conversió hauran d'estar actualitzades per a poder treballar amb elles, independement de si la lectura ha estat acabada o no.
Stakeholder	Estadístics.
Prioritat	Mitjana

Taula 13.9: RNF - P01: La inserció no ha d'aturar el treball

Identificador	RNF-P02
Nom	La interacció amb el sistema ha de ser simple.
Característiques	La corba d'aprenentatge de cara a l'usuari ha de ser el màxim de curta possible.
Descripció	El sistema ha de resultar fàcil d'usar i utilitzar estructures i modes de funcionament semblants als softwares ja en ús de l'empresa. Això és degut al fet que, quant més semblant a un software ja en ús, menys resistència al canvi existirà.
Stakeholder	Estadístics.
Prioritat	Baixa

Taula 13.10: RNF - P02: Interacció amb el sistema simple.

13.4 Requisits no funcionals: Limitacions disseny

Identificador	RNF01
Nom	El sistema ha de ser integrable en els softwares R, Excel i Tableau.
Característiques	El sistema ha de ser integrable en els softwares R, Excel i Tableau per a l'obtenció de datasets.
Descripció	El sistema ha d'estar muntat sobre una plataforma directament accessible desde R, Excel i Tableau.
Stakeholder	Equip directiu, analista programador i estadístics.
Prioritat	Alta

Taula 13.11: RNF - LD01: Integrable en els softwares R, Excel i Tableau

Identificador	RNF02
Nom	El sistema ha de ser compatible amb Java i VB0.
Característiques	El sistema ha de ser compatible amb Java i VB0 per a la seva futura integració en softwares de l'empresa.
Descripció	El sistema ha de ser compatible amb Java i VB0 per a la seva futura integració en softwares de l'empresa. Qualsevol desenvolupament realitzat per tant, haurà de ser tractat amb un d'aquest dos llenguatges, preferiblement VB0.
Stakeholder	Equip directiu, analista programador.
Prioritat	Alta

Taula 13.12: RNF - LD02: Compatible amb Java i/o VB0.

Identificador	RNF03
Nom	El sistema ha de ser aplicable amb el hardware existent.
Característiques	El sistema ha de ser integrable amb el hardware existent a l'empresa.
Descripció	El sistema ha de ser integrable amb el hardware existent a l'empresa. Cal tenir en compte que els servidors existents poden ser estesos, pel fet que la capacitat del rack encara no ha estat completament usada i per tant, es pot afegir memòria i capacitat de processament en un nou servidor virtual.
Stakeholder	Equip directiu.
Prioritat	Alta

Taula 13.13: RNF - LD03: Aplicable amb el hardware existent.

Identificador	RNF04
Nom	S'ha d'usar un sistema de bases de dades relacional.
Característiques	El sistema ha de ser dissenyat per a usar un sistema de bases de dades relacional.
Descripció	El sistema ha d'estar preparat per usar una base de dades relacional. És preferible l'ús de tecnologies ja conegudes com MSSQL, MySql, etc.
Stakeholder	Equip directiu, Analista programador.
Prioritat	Alta

Taula 13.14: RNF - LD04: S'ha d'usar un sistema de bases de dades relacional.

Identificador	RNF-LD05
Nom	L'aplicació haurà de seguir els patrons de marca de l'empresa.
Característiques	La paleta de colors i el disseny de l'aplicació haurà de seguir els patrons de disseny d'apps de l'empresa.
Descripció	El sistema d'interfícies a dissenyar haurà de seguir els patrons de disseny que s'han generat per la marca Ikimedia Communications S.L. Ha de seguir el patró d'imatge de la marca.
Stakeholder	Equip Directiu.
Prioritat	Mitjana

Taula 13.15: RNF - LD05: L'aplicació haurà de seguir els patrons de disseny de l'empresa.

13.5 Requisits no funcionals: Atributs qualitat

Identificador	RNF-AQ01
Nom	El sistema ha de ser fàcil de mantenir
Característiques	El sistema ha de ser el màxim de fàcil de mantenir
Descripció	El sistema plantejat haurà de ser gairebé automantingut, no poden existir tasques diàries a realitzar a part de l'obtenció de les dades que s'han d'introduir manualment.
Stakeholder	Analista Programador.
Prioritat	Mitjana

Taula 13.16: RNF - AQ01: Fàcil de mantenir

Identificador	RNF-AQ02
Nom	El sistema ha de mantenir la privacitat de les dades dels diferents usuaris
Característiques	El sistema ha de mantenir privacitat en les dades per llei i a més a més ha d'assegurar que no donarà descripcions a tothom de les accions que cada treballador realitza.
Descripció	La solució proposada haurà d'oferir un sistema de permisos de visualització de les dades dels usuaris per impedir que les dades d'un usuari quedin públiques pels demés sinò tansols per certs usuaris amb drets de controlar el treball realitzat.
Stakeholder	-. Equip directiu.
Prioritat	Mitjana

Taula 13.17: RNF - AQ02: Mantenir la privacitat de les dades dels diferents usuaris

13.6 Traducció a casos d'ús

Per tal d'identificar més concretament els casos d'ús referenciats pels requisits, s'ha decidit presentar el gràfic de casos d'ús que ens permetrà entendre més especificament les funcionalitats que demana el sistema. Aquests casos d'ús per tant, tansols referencien explícitament els requisits funcionals tot i que alguns poden referenciar també de no funcionals de manera indirecta.

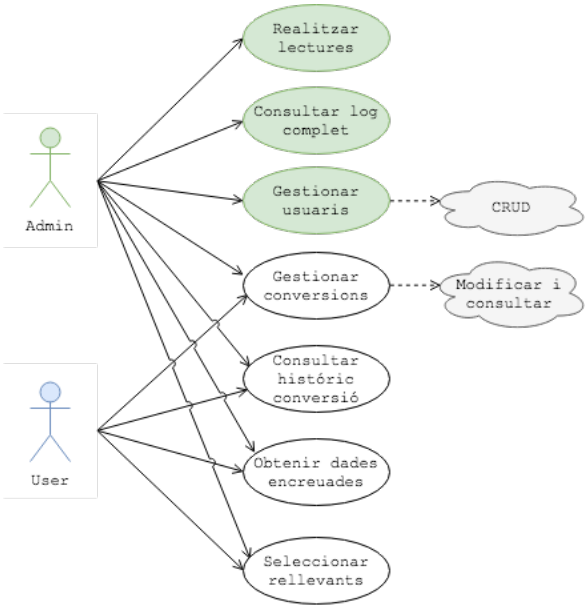


Figura 13.2: Diagrama de casos d'ús

Veiem doncs en el gràfic que existeixen dos tipus d'usuaris: l'usuari de tipus administrador i l'usuari de tipus bàsic. Veiem també que l'usuari administrador realitza totes les funcions que pot fer l'usuari bàsic i per tant, suposarem que:

Figura 13.3: Diagrama de definició dels usuaris

D'aquesta manera (Figura 13.3) podem assegurar que, si un usuari pot complir una funcionalitat X, un administrador podrà complir-la també, ja que, tindrà accés a aquesta. Passem ara a realitzar la descripció de cada un dels UC per tal d'intentificar-los i saber què es demana de cada un d'ells.

ID	Nom	Descripció
UC00	Log in or out	Qualsevol usuari ha de ser capaç de autenticar-se o tancar sessió en el sistema.
UC01	Realitzar lectures	Aquest cas d'ús busca oferir la lectura de les dades a partir d'un fitxer de tipus .txt (aquest format ve donat de l'estudi de fonts de dades realitzat, on s'ha vist que gairebé totes són extretes en aquest format o en csv). Aquesta tasca tan sols serà executada per l'administrador.
UC02	Consultar log complet	Un administrador ha de poder consultar el log complet de les accions realitzades sobre la base de dades i tenir la capacitat de filtrar-les fàcilment per identificar possibles errors o canvis injustificats.
UC03	Gestionar usuaris	Un administrador ha de ser capaç de crear, modificar, eliminar i consultar els usuaris existents a la base de dades.
UC04	Gestionar conversions	Un usuari ha de poder modificar i consultar les conversions existents per a cada un dels clients i fonts de dades. A més, ha de poder filtrar per veure casos no recatalogats importants.
UC05	Consultar històric de conversió	Un usuari ha de ser capaç de, a partir de la consulta d'una conversió, obtenir informació sobre l'històric d'aquesta conversió.
UC06	Obtenir dades encreuades	Un usuari ha de ser capaç de realitzar una consulta SQL amb les dades encreuades des de qualsevol dels softwares seleccionats. Per fer-ho, com ja s'ha explicat a l'anàlisi de requisits, l'usuari no tindrà interfície, sinó que usarà SQL.

Taula 13.18: Descripció dels casos d'ús existents

14

Selecció del Software

Tot i que inicialment aquesta selecció estava plantejada per a ser feta una vegada introduït el disseny inicial de la solució, a causa de l'estudi de context, ens n'hem adonat que resulta molt més coherent seleccionar amb anterioritat el software, ja que, molts dels requisits o de les condicions del context es veuran satisfetes de manera més completa si es fixen determinats softwares.

Per tant, procedim a realitzar la selecció del software en base a l'estudi de context realitzat i intentant que aquest software sigui el màxim d'ajustat. Buscant així que existeixi el mínim de resistència al canvi possible per part dels diferents stakeholders. Tanmateix, també es buscarà usar softwares que no modifiquin l'infraestructura informàtica existent i en cas de ser softwares nous, que puguin ser usats de manera gratuïta o amb un cost molt reduït. Per a la realització d'aquest apartat pretenem dividir la selecció en els següents punts:

- *Base de dades:* Es refereix tant a la selecció al sistema de gestió de bases de dades relacional com al client de base de dades proposat (és proposat perquè, cada usuari podria usar un client diferent i el funcionament del sistema no hauria de modificar-se).
- *Llenguatge de programació:* Es refereix a la selecció del llenguatge de programació i les diferents llibreries i plug-ins a usar.
- *Software per a la programació (IDE):* Referit entorn de desenvolupament integrat que s'usarà. També s'inclou en aquesta secció softwares rellevants pel desenvolupament de la solució.
- *Softwares als que s'assegura l'integració:* Conjunt de softwares als quals la solució haurà d'assegurar la integració i justificació d'aquesta integració.
- *Altres:* D'altes softwares a usar durant el desenvolupament que no són de gran rellevància de cara al projecte (i.e. editor d'imatges).

Comencem doncs la selecció del software a tenir en compte de cara a la solució proposada pel punt de tria de la base de dades.

14.1 Base de dades

14.1.1 Sistema de gestió de bases de dades relacionals

En un primer moment el sistema de gestió de bases de dades relacionals proposat va ser MySQL, la plataforma de codi obert de gestió de bases de dades més gran del món. Aquesta selecció s'havia realitzat gràcies a la gran popularitat en l'àmbit d'aquest gestor i a què, la persona que s'haurà d'encarregar del desenvolupament del projecte, és a dir l'Enginyer Informàtic [jo], tenia experiència prèvia en MySQL i en les llibreries vinculades a aquest gestor de base de dades facilitant així doncs la realització del projecte.

Tot i això, finalment no s'ha acabat seleccionant MySQL, sinó que s'ha tendit per usar Microsoft SQL Server (MSSQL). Aquesta decisió no es realitza de forma arbitrària sinó que es pren a partir de les següents informacions recollides en l'estudi de context realitzat:

- *El software usat és microsoft.* Del punt d'estudi dels softwares n'extraiem la conclusió de què gran part dels softwares usats són Microsoft Office, a més a més, com s'ha vist al punt de l'estudi del hardware els sistemes operatius de la majoria d'usuaris són Windows (Microsoft) fet que, fa que tots els altres softwares existents assegurin una integració completa amb Microsoft. Per tant, un dels motius de la selecció d'aquest gestor és per assegurar-ne la integració amb el màxim de softwares possibles encaixant així amb part del requisit RNF-01.

- *En el hardware actual ja existeix un servidor dedicat a aquest gestor.* Durant l'estudi del hardware realitzat s'ha descobert que existeix un servidor virtual anomenat 'sql'. Aquest servidor és un Microsoft SQL Server i per tant, amb la selecció d'aquest gestor ens evitarem la creació d'un nou servidor virtual al rack, reduint així la feina a realitzar per part dels gestors del servidor (fet citat en l'estudi inicial dels stakeholders com a objectiu). Així doncs amb aquesta decisió ens ajustem al requisit RNF-LD03.

- *La persona que haurà de gestionar el sistema en un futur (Analista programador) té experiència amb MSSQL*. Durant l'estudi del stakeholder 'Analista programador' ens n'hem adonat que resulta ser una persona amb una formació de nivell bàsic però amb un nivell d'experiència molt elevat. Aquest stakeholder ha treballat amb MSSQL i per tant, pot tant resultar d'ajuda en el desenvolupament, com facilitar el manteniment de la base de dades un cop implementada. En resultar a més una figura de poder, s'ha d'aconseguir de reduir amb màxim la seva resistència al canvi i es creu que facilitant un gestor de base de dades en el que ja tingui experiència aquesta serà reduïda de manera representativa. Aquest motiu per tant s'ajustarà als requisits RNF-AQ01 i RNF-P02.

- *Gestió d'usuaris.* La connexió a aquesta base de dades es pot realitzar mitjançant Windows Authentication. Aquest fet ens permetrà no haver de gestionar els usuaris en primera instància i és podran usar els usuaris autenticats en el sistema. D'aquesta manera reduïm el nombre de comptes que han de memoritzar els treballadors (veient la llista de softwares estudiats en el punt d'anàlisi d'aquests ens n'adonem que ja n'han de recordar mínim 4 de diferents) i alhora facilitem la interacció reduint els passos per accedir a l'aplicació, ens estalviem el log in. Per tant complim amb els requisits RNF-P02 i RNF-AQ02.

- *Funcionalitats adequades pel cas d'estudi.* Trobem també que aquest gestor de bases de dades relacionals ofereix les mateixes funcionalitats que MySQL i que per tant és adient per l'estudi. Cal recordar que una limitació de cara a aquest projecte era que fos una base de dades relacional (RNF-LD04) fet que MSSQL assegura.

- *Documentació extensiva*. Finalment, i un cop decidit que era la millor opció pel cas d'estudi. S'ha buscat si existia documentació suficient per a poder aprendre a usar-lo de manera ràpida. També s'ha fet una cerca ràpida de les possibilitats de solució de problemes de la comunitat d'usuaris d'aquest sistema i s'ha detectat que era molt extensiu en ambdós casos i que, per tant, és factible l'aprenentatge per part del desenvolupador d'aquest nou entorn de manera ràpida i es valora que aporta suficient valor per a considerar-ne l'aprenentatge.

Concloem doncs que s'usarà el gestor de base de dades relacionals MSSQL pels motius prèviament exposats.

14.1.2 Client de bases de dades

Per a la interacció amb la base de dades s'han proposat dos softwares que són gratuïts i permeten la gestió mitjançant un front-end amb la base de dades, evitant així haver de realitzar-ho tot via consulta (complint doncs amb el RNF-P02 i el RNF-AQ01). S'ha decidit seleccionar els softwares en dependència del sistema operatiu usat per l'usuari, tot i que, en cas de voler, els usuaris de sistemes Windows podrien usar qualsevol dels dos. Els softwares de client de base de dades elegits han estat Microsoft SQL Server Management Studio i TeamSQL.

Pel que a a Microsoft SQL Server Management Studio s'ha triat a causa d'una recomanació del 'Analista progamador' que ha treballat durant molt de temps amb aquest software i el col·loca com una de les millors plataformes per a la gestió de bases de dades. Ens informa que aporta maneres visuals d'interaccionar amb la base de dades arribant a separar el codi de la gestió, és a dir, inclou funcionalitat com ara la creació de taules des de front-end sense necessitat d'escriure codi SQL. A més a més, ofereix un sistema d'estudi de les diferents queries que es llencen a la base de dades que notifica en cas de trobar possibles optimitzacions de les consultes. Per tant, amb aquest software ens assegurem que els usuaris de manteniment podran ser persones sense un coneixement molt elevat de SQL i que a més a més, novament tornarem a posar-nos prop de l'analista programador, reduint la resistència al canvi que s'espera que ofereixi el mateix. Així doncs, la selecció d'aquest software ens permet apropar-nos al compliment dels requisits RNF-02 i RNF-AQ01. Com a contra punt, aquest software tan sols serà usable des de sistemes Windows ja que, no està preparat per a MacOs. Tot i això, aprofitant els servidors virtuals un usuari de Mac podria usar-lo amb la connexió a l'escriptori remot.

En qualsevol cas, si un usuari de MacOs no vol usar la màquina virtual s'ha trobat una opció extra que permet accedir des de qualsevol sistema operatiu. Aquesta és TeamSQL i permet ser usat des de qualsevol dels dos sistemes operatius però ofereix un sistema de gestió basat plenament en codi i per tant no resulta tan fàcil d'usar com l'anteriorment anomenat. Tot i això també té els seus punts forts com ara la possibilitat d'integració amb google i la gestió d'equips de treball. També ofereix la graficació directe de les dades emmagatzemades.

Justifiquem doncs que les dues opcions triades són Microsoft SQL Server Management Studio i TeamSQL, essent la primera la prioritària i la segona la secundària. El fet de tenir aquestes dues opcions i no tan sols la primera ens permet assegurar que el requisit RNF-LD03 es compleix, ja que, tots els usuaris podran accedir-hi sense haver de canviar de software i perquè, en cas de tan sols fer servir la primera opció, mitjançant el servidor amb escriptori remot es podria accedir al client.

14.2 Llenguatge de programació

Finalment el llenguatge de programació elegit ha estat Java. Aquesta decisió s'ha pres condicionada l'experiència de l'Enginyer Informàtic amb aquest llenguatge. A més a més, la majoria de patrons de disseny existents són molt fàcils d'implementar en Java, ja que, a diferència d'altres llenguatges està molt preparat per treballar amb conceptes com Interfícies, Singletons, Herència, etc. Fet que farà més simple el procés de desenvolupament. A més a més, s'està usant un llibre anomenat Head First Desing Patters de Eric Freeman i Elisabeth Robson que ajuda a comprendre els diferents patrons i està exemplificat en Java. Amb la selecció d'aquest llenguatge doncs es compleix el requisit RNF-LD02.

També s'ha decidit usar un plug-in afegit a Java que permet la realització del front-end de tipus Rich Internet Application (RIA). Aquesta decisió s'ha pres a causa de l'interès a tenir un front-end el màxim d'user friendly (RNF-P02) i per poder seguir amb l'estètica oferta per l'empresa de cara al desenvolupament del front-end (RNF-LD05). Aquest llenguatge permet, a partir d'un sistema semblant al xml (.fxml), generar la definició de les diferents pantalles de manera que siguin 'responsive', fet que la fa més usable (RNF-P02). Aquest plug-in no modifica el IDE i per tant, segueix complint amb el requisit RNF-LD02.

14.3 Software per a la programació

Pel que fa al software usat per a la programació del llenguatge Java s'usarà el software més conegut per aquest llenguatge, és a dir, Eclispe. En aquest cas s'usarà Eclipse Neon, essent aquesta l'última versió. Aquest software és usable des de qualsevol dels sistemes operatius existents a l'empresa i s'instal·larà a la màquina virtual per assegurar que s'hi pot accedir des dels ordinadors de fora de l'empresa. Donant la possibilitat de solucionar minor bugs del sistema en cas d'urgència sense necessitat d'estar a l'oficina. Amb aquesta última característica afegim valor al requisit RNF-AQ01 i RNF-LD04

A més a més d'aquest software, s'usarà Atom com a editor de text bàsic per a casos especials en els quals és vulgui modificar ràpidament un arxiu sense necessitat d'obrir l'entorn de programació. Atom és un software lliure i usable des de qualsevol sistema operatiu que permet modificar text oferint llibreries que afegeixen highlights (colors) i autofill (auto-completació) facilitant així la programació. Per tant, aquest és un software que ens permetrà implementar les diferents pantalles i comprovar-les sense necessitat d'usar directament el .fxml. En general, aquest software serà usat per generar la plantilla de pantalla i posteriorment mitjançant canvis en el codi s'aconseguirà el resultat final desitjat. Amb aquest software facilitem la feina del desenvolupador.

14.4 Softwares als que s'assegura l'integració

Acabem amb la selecció assegurant la integració del sistema amb un seguit de softwares i per fer-ho és dona el nom del software i una demostració de la seva possible integració amb MSSQL, que és la base del sistema. Recordem que els softwares a integrar van relacionats amb l'obtenció de dades i que l'únic cas en el qual existeix possibilitat de modificació és en el cas del software desenvolupat i per tant, l'únic que ha de poder oferir opcions de modificació sobre MSSQL ha de ser el llenguatge Java. Comencem doncs l'enumeració d'aquests softwares:

- *Tableau:* De Tableau s'ha trobat una explicació dels mateixos proveïdors del servei on expliquen com realitzar la integració de Tableau i MSSQL.
- *R:* De R també trobem una explicació feta per R Studio de com connectar amb MSSQL. En aquest cas però es necessitarà usar un plugin anomenat ODBC que ens permetrà la connexió entre les dues plataformes.
- *Java:* A Java existeix una llibreria anomenada Connection que ofereix la possibilitat de conectar-se a qualsevol tipus de base de dades i interactuar via codi SQL amb ella. D'aquesta mateixa manera existeixen llibreries com ResultSet, PreparedStatement, etc. que permeten gestionar tot tipus de crides a la base de dades. Podem trobar més documentació d'aquest fet aquí
- *Excel:* Excel ofereix un conjunt de funcions dedicades a la connexió amb fonts dades i per tant, també podrà connectar-se com podem veure en aquesta web.

Podem assegurar doncs que tots els softwares poden ser integrats amb MSSQL i que per tant estem complint els requisits RNF-LD01 i RNF-LD02 dedicats a la integració amb els diferents softwares.

14.5 Altres

Acabem anomenant d'altres softwares que s'usaran per a la realització del projecte tot i que no resulten especialment rellevants de cara als requisits. Aquest són: Photoshop (per l'obtenció de les imatges usades en el frontend), LaTexIt (per generar la documentació) i Mendeley (per gestionar les referències).

15

Selecció del Hardware

Com ja s'ha citat anteriorment, ha existit un canvi en la planificació passant a realitzar-se primer els punts referents a la selecció de software i de hardware abans que la proposta de solució per a poder tenir més en compte molts factors que s'han descobert durant l'estudi de context que altrament podrien no ser considerats. Així doncs comencem la selecció del hardware i les peculiaritats d'aquest.

Ja s'ha parlat que és pretén muntar el servidor MSSQL sobre el servidor virtual ja existent en el rack per reduir la càrrega a donar als gestors del servidor. D'aquesta manera fem que aquest projecte encaixi més amb els objectius identificats dels mateixos en l'estudi inicial i complet de stakeholders. A més a més d'aquesta manera el desenvolupament es pot realitzar sense haver d'esperar que és crei un nou servidor virtual al rack, tasca que sol tenir una demora d'uns 5 dies laborables (reduint així doncs el temps estimat d'implementació d'un prototip a una sola setmana en cas de ser possible). En tot cas aquest servidor seleccionat podrà ser ampliat tant en memòria com en potencial de càlcul en qualsevol moment, assegurant així l'escalabilitat del sistema en qüestió.

Apart d'aquests dos servidors virtuals (sql i dc), també es considera necessari l'ús del servidor 'back' amb l'objectiu d'oferir un servidor de còpia de les dades de la base de dades per si en algun moment s'ha de tirar enrere un conjunt d'accions realitzades. Aquesta còpia de dades actualment ja es realitza de manera automàtica per a tots els servidors virtuals existents i per tant, no resultarà necessari aplicar cap canvi per mantenir-ho a l'estar fent servir el servidor sql ja inicialitzat en el rack.

A més a més, també s'haurà de tenir en compte el servidor de connexió externa amb escriptori remot (gen), ja que, s'usarà com a mètode d'accés extern a l'empresa per part dels treballadors.

D'altres components importants que s'hauran de mantenir seran, el router per realitzar el control d'accés, els hubs de les dues plantes (ignorant el de convidats, ja que no té accessos als servidors) i el proxy amb firewall. Finalment també serà necessari tenir en consideració que existeixen els ordinadors personals i els d'empresa que poden tenir tant un sistema operatiu MacOS com Windows.

Així doncs el software seleccionat resulta ser el mostrat a la figura X. On és mostra un subconjunt del hardware estudiat al punt d'estudi del hardware existent de l'estudi de context realitzat.

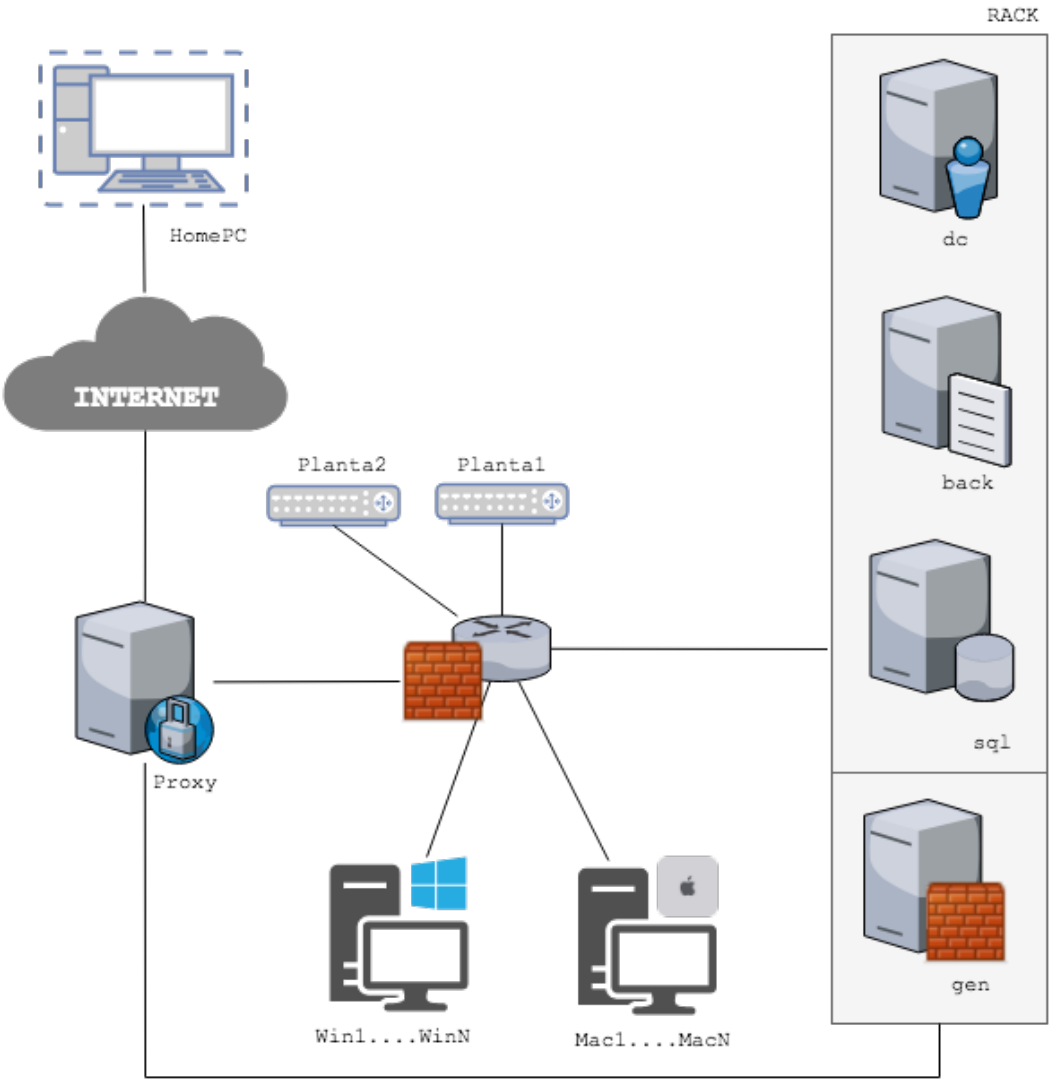


Figura 15.1: Hardware seleccionat

Recordem també que aquest servidor resulta tan sols accessible des de la xarxa local de l'empresa assegurant la protecció de dades demanada en el requist RNF-AQ02. L'únic mètode d'accés al servidor de manera externa és mitjançant una connexió directa al servidor d'escriptori remot gen. D'aquesta manera podem assegurar que el sistema serà també accessible des de casa dels treballadors sense deixar de funcionar tan sols dins de la xarxa local. Per clarificar l'explicació es presenta una versió del gràfic anterior gràfic (figura X) on s'explica amb fletxes verdes una connexió interna i amb taronja una connexió externa per demostrar que l'accessibilitat estarà assegurada.

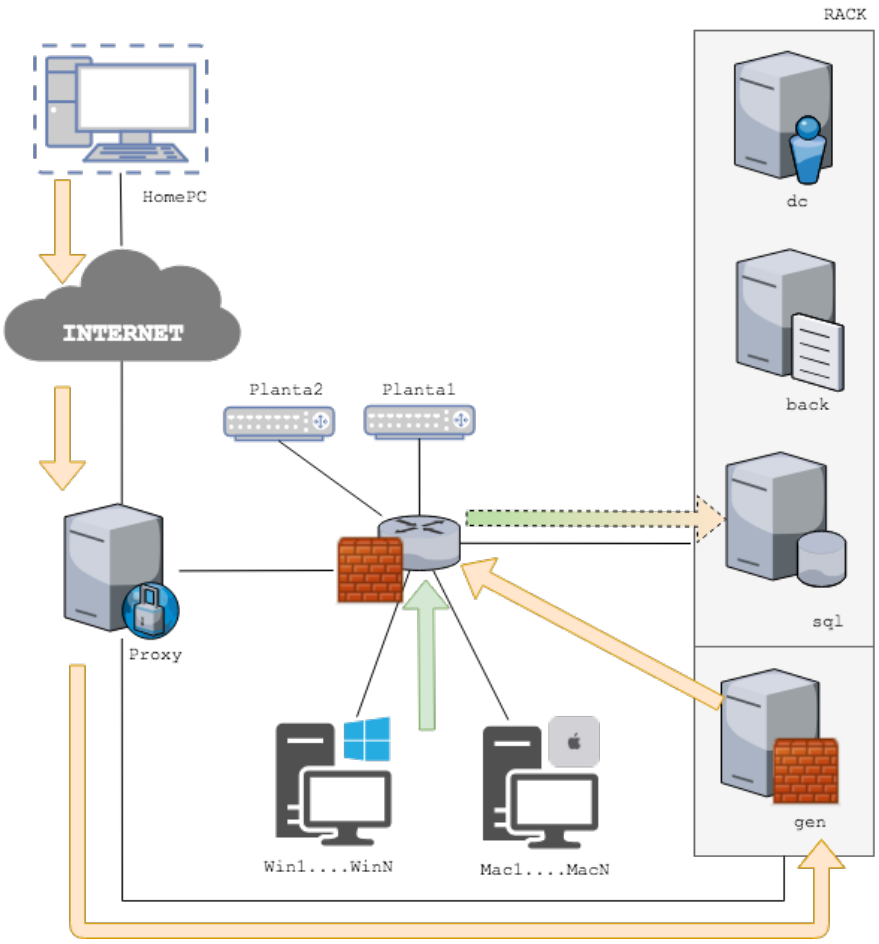


Figura 15.2: Seguretat conexions

Veiem doncs que pel que fa a les connexions internes, els diferents ordinadors es connecten al router, ja sigui mitjançant la connexió directa al router o la connexió mitjançant els hubs que redirigeixen al router. El router comprova que les connexions són dins del rang IP esperat i en cas afirmatiu, envia al 'rack' en cerca del servidor virtual demanat. D'aquesta manera s'estableix una connexió que tan sols podrà ser local i en unes IPs donades, ja que, es realitza un filtre al router que no permetrà connexions externes o de dispositius no desitjats, per exemple del hub 'invitados'.

Parlant ara de la connexió externa, ens trobem que l'ordinador personal de cada treballador enviarà un senyal a internet que passarà directe al proxy redirigint-la mitjançant el router (no indicat en la imatge per simplificar el gràfic) al servidor 'gen' on establirà una connexió d'escriptori remot. Un cop iniciada aquesta connexió d'escriptori remot, l'usuari podrà interaccionar igual que si tingués un ordinador a l'empresa, ja que, qualsevol petició feta des de l'escriptori remot suposarà l'enviament al router d'una connexió amb una ip local vàlida que permetrà connectar-se als diferents servidors virtuals.

Queda doncs exposada la selecció de hardware realitzada i es pot comprovar que el hardware triat és un subconjunt del hardware existent a l'empresa i que per tant s'ha complert el requisit RNF-LD03, ja que, no s'ha necessitat incrementar en hardware de cap tipus. Val a dir però que per motius d'escalabilitat, sí que es contempla la futura millora del servidor virtual 'sql' en cas de tirar-se endavant el projecte incrementant-ne tant l'espai de memòria com el potencial de processament.

16

Proposta de solució

Una vegada establertes les condicions que s'han de complir per a obtenir una solució vàlida, es procedeix a realitzar una proposta de solució que encaixi amb el context estudiat. Per tant, aquesta solució, haurà tant de tenir en compte tot l'estudiat durant l'estudi de context, tots els requisits plantejats hauran de ser complets i els condicionants de software i hardware afegits també hauran de ser tinguts en compte. Aquest punt resulta doncs una integració de tot l'informe realitzat fins ara que acabarà amb el plantejament d'una solució vàlida.

Per tal d'assegurar que aquesta relació entre la proposta i els conceptes estudiats es compleix es pretén usar el següent sistema de verificació basat en tres preguntes (al final de la proposta, tot i que, durant el transcurs de l'exposició de la proposta també es vincularan les decisions amb els conceptes dels punts anteriors), una per a cada paquet del treball:

1. Estudi de Context (punts X): S'han usat els conceptes estudiats durant l'estudi de context?
2. Anàlisi de requisits (punt X): Es compleixen tots els requisits?
3. Selecció de hardware i software (punt X): Es pot implementar amb els softwares i hardwares usats?

L'estructura d'aquesta proposta també està determinada, es començarà donant un diagrama de classes que permeti encabir totes les dades extretes fonts de dades estudiades, posteriorment es realitzarà la traducció d'aquest a diagrama de modelatge de base de dades amb l'objectiu assegurar que això podrà realitzar-se en una base de dades relacional MSSQL.

Un cop realitzats els passos anteriors es donarà per finalitzada l'exposició del back-end necessari i es passarà a descriure una interfície gràfica que encaixi amb les peculiaritats del cas d'estudi que s'està portant a terme. Per tal de fer-ho s'usaran dos artefactes de l'enginyeria del software anomenats mock-ups (vistes del programa) i diagrames de navegació, d'aquesta manera es pretén forçar tant que la interfície proposada segueixi els requisits establerts, com que sigui realitzable amb el back-end que prèviament haurà estat dissenyat.

Finalment i tancant la solució es donarà una conclusió on es responguin les tres preguntes prèviament exposades i s'asseguri de manera genèrica que el funcionament de la proposta està assegurat.

16.1 Disseny d'un diagrama de classes inclosiu

En aquest punt es pretén aconseguir realitzar un diagrama de classes que permeti incloure totes les fonts de dades estudiades i considerades, és a dir, les dades globals (KM, INV i IOPE) i les dades específiques de client, per tal de poder-ho traduir posteriorment a un diagrama de modelatge de base de dades que sigui introduïble a la base de dades MSSQL.

Amb l'objectiu d'acabar obtenint un diagrama complet s'ha dividit aquesta tasca en diferents apartats que busquen acostar-nos de manera continuada cap a un diagrama de classes inclusiu mitjançant l'addicció separada de conceptes observats durant l'estudi de context i l'anàlisi de requisits. Aquest procés es basarà doncs en els 3 passos següents:

1. *Integració de les fonts globals:* Es buscarà trobar un diagrama (independent dels valors de les dades, és a dir, tan sols tenint en compte les definicions de les mateixes) que permeti encabir totes les fonts globals a estudiar (KM, INV, IOPE).

2. *Introducció del problema de valor vs. definició:* En aquest punt es pretén trobar una proposta que permeti solucionar el problema existent que entre diferents fonts de dades els camps que són definits de manera molt similar o igual tinguin valors que expressin el mateix però amb diferent format. Per exemple, que el camp sector en una font sigui 'Pelo' i en un altre 'pelos', queda clar que ambdues busquen expressar un mateix sector però amb valors diferents fet que fa que el seu encreuament no pugui realitzar-se automàticament.

3. *Integració amb les dades extres de client:* Un cop totalment integrades les dades globals es buscarà una solució que permeti integrar aquestes dades globals amb les dades extres proporcionades pel client per a tots els casos. Recordem que en aquest cas, no s'ha de realitzar el diagrama de classes de cada un dels clients sinó que s'ha de donar una proposta de funcionament que ofereixi un mètode d'encreuament vàlid per a tots els clients.

Una vegada realitzats aquests punts es creu que s'obtindrà un diagrama de classes inclusiu per a totes les fonts existents però, tot i això faltarà per realitzar una última consideració important de cara al cas d'estudi al qual s'aplica per tal de poder ajustar aquest esquema a la realitat en la qual ens trobem. Aquesta serà la següent:

- Cada client pot tenir una visió de mercat diferent i considerar de manera diferent els encreuaments? És a dir, per exemple, pot passar que un client entengui el sector 'Peluqueria' com a 'Pelo' i un altre l'entengui com a dos sectors diferenciats?

Com ja s'ha dit anteriorment probablement sense realitzar aquesta consideració ja s'obtindria un diagrama de classes que inclogués totes les fonts, el problema és que, molt possiblement no seria aplicable al cas d'estudi, sinó que tan sols seria aplicable de cara a un sol client. Cal doncs reflexionar sobre si existeix una necessitat real d'introduir la figura de client per assegurar que tota la informació pot ser tractada amb el diagrama de classes ofert o si no resulta rellevant de cara a la realització d'aquest projecte.

Iniciem doncs aquest procediment d'estudi realitzant la integració de les fonts globals un cop explicat el flux de treball que se seguirà per a aquest apartat.

16.1.1 Integració de les fonts globals

Comencem doncs realitzant aquesta primera integració de fonts globals com a primer pas cap a la integració inclusiva final. Per fer-ho és tindran en compte les fonts de dades globals estudiades durant l'estudi de les fonts de dades, és a dir, KM, INV i IOPE. Cal recordar que aquesta integració tindrà en compte tan sols la definició dels camps, és a dir, vincularà els conceptes de les diferents fonts sense donar importància al valor que aquest agafi (recordem l'exemple de 'Pelo', 'pelos' prèviament explicat). Aquest fet que no serà tractat en aquest punt, pot veure's estudiat en el punt d'Introducció del problema de valor vs. definició. Així doncs, de cara a aquest estudi tan sols es tindran en compte els punts d'estudi de les fonts de dades globals i especialment, dins d'aquests apartats, la definició UML realitzada i l'exposició de les metadades (tan sols la definició de cada camp).

Primerament i aprofitant l'estudi de les fonts de dades realitzat exposem els diferents diagrames de classes que s'han realitzat per a cada una de les fonts globals amb la intenció de trobar similituds que ens permetin integrar-les en un diagrama comú (Figures 16.1, 16.2, 16.3).

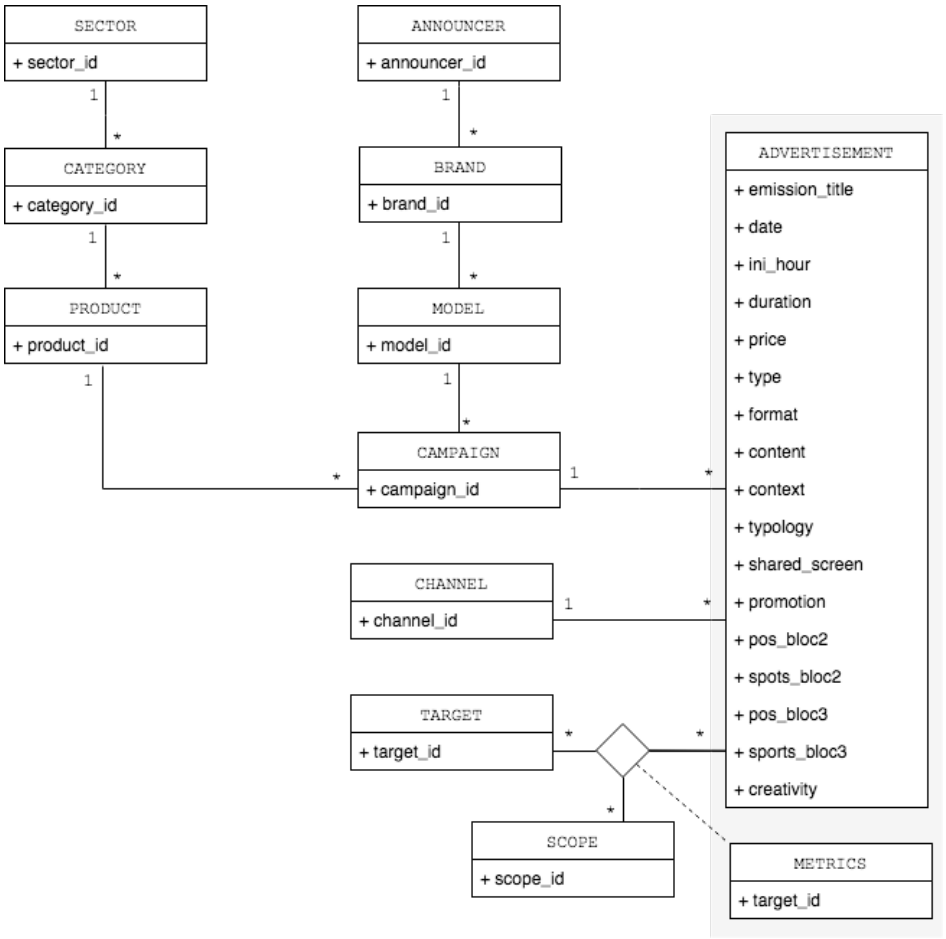


Figura 16.1: Diagrama de classes UML Kantar Media [km]

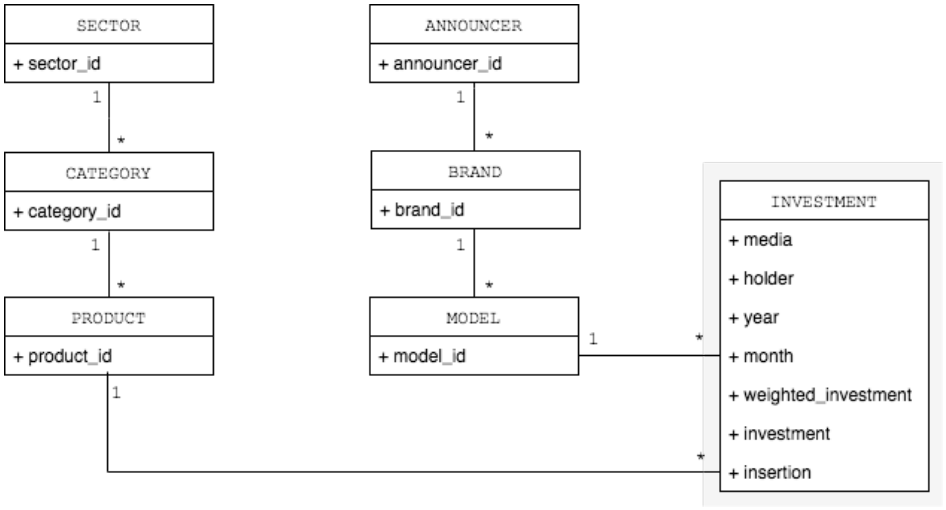


Figura 16.2: Diagrama de classes UML Info Adex [inv]

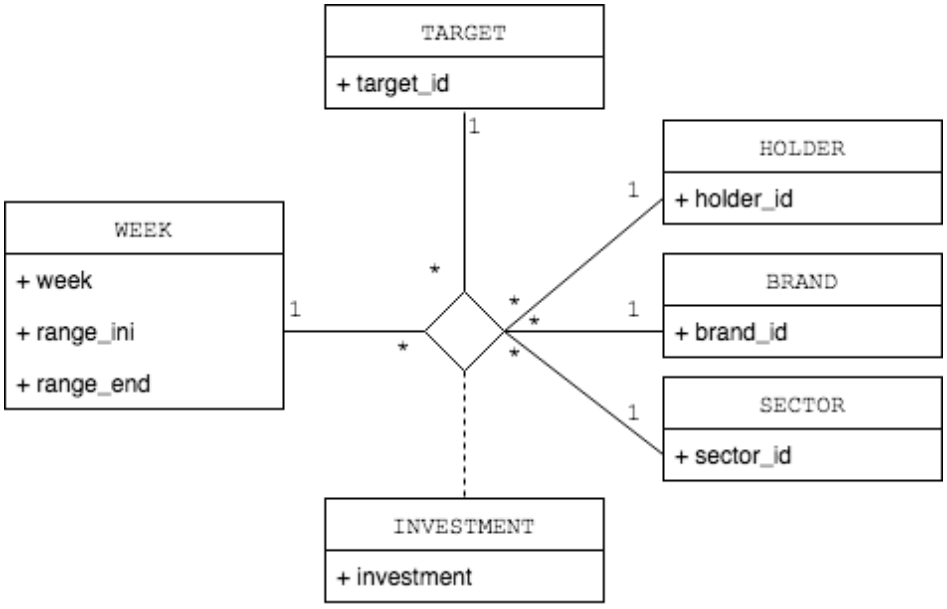


Figura 16.3: Diagrama de classes UML Kantar TNS [iope]

Amb la presentació conjunta d'aquests diagrames de classes descobrim que existeixen atributs i classes que es repeteixen. Aquest fet però no assegura que siguin el mateix quant a definició, per exemple, imaginem que tenim dues classes una classe Persona i una classe Ciutat, en cas que les dues continguin l'atribut Nom no voldria dir que Persona.nom === Ciutat.nom, ja que, la seva respectiva definició no seria la mateixa i per tant no serien equiparables, en aquest cas la definició d'una seria nom de la persona i l'altre nom de la ciutat i per tant, com que les definicions no resulten iguals aquests camps no representarien el mateix.

Observem doncs que les classes/atributs que es repeteixen són sector, announcer, category, brand, product, model i target. Aquestes classes però, no apareixen per a totes les fonts de dades. Veiem també que, la majoria d'aquestes similituds són donades pels camps d'agrupació de cada una de les fonts, és a dir, els camps descriptius que en una base de dades podrien ser entesos com a part de les claus primàries o secundàries. És observable també que tot i compartir aquestes classes, no mantenen les mateixes relacions i per tant s'haurà de realitzar un plantejament de com mantenir aquestes o si mantenir-les.

Un cop vist que aquestes classes/atributs poden ser el principal punt de vincle entre les fonts i explicada la problemàtica existent amb la definició dels camps, cal comprovar si realment les classes tenen el mateix significat entre les diferents fonts per poder integrar-les en una sola classe. Per tal de fer aquest pas es presenten les files de les taules de meta dades estudiades durant l'anàlisi de context per tal de veure si les definicions encaixen i si per tant, aquestes classes són integrables entre si. Podem veure-ho a les taules 16.1, 16.2 i 16.3.

Atribut	Tipus	Descripció	Exemple
Sector	String	Sector en el qual es troba el producte anunciat	BELLEZA e HIGIENE
Categoria	String	Categoria que te dins del sector assignat	Productos CABELLO
Producto	String	Línia de productes dins de la categoria assignada	Fijadores y Moldeadores
Anunciante	String	Anunciant que paga per l'anunci del producte	PROCTER and GAMBLE ESPAÑA, S.A.
Marca	String	Marca a la que representa l'anunciant	PANTENE PRO-V
Modelo	String	Model que pretén promocionar la marca amb l'anunci	ESPUMA RIZOS
Campaña	String	Campanya publicitaria a la qual pertany l'anunci, sovint combinació de marca + model	PANTENE PRO-V/ESPUMA RIZOS

Taula 16.1: Kantar Media metadata

Atribut	Tipus	Descripció	Exemple
Sector	String	Especifíca el sector en el que és troba el producte anunciat	BELLEZA E HIGIENE
Categoria	String	Especifíca la categoria dins del sector del producte anunciat	PRODUCTO CABELLO
Producto	String	Especifíca la linea de productes dins de la categoria del producte anunciat	ACONDICIONADORES Y SUAVIZANTES
Anunciante	String	Especifíca l'anunciant que paga per l'anunci del producte	PROCTER and GAMBLE ESPAÑA, S.A.
Marca	String	Especifíca la sub-marca a la que representa l'anunciant (Gairebé sempre resulta ser Marca == Marca Directa)	PANTENE
Modelo	String	Especifíca el model que preten promocionar la marca amb l'anunci	PRO V REPA.PRO.

Taula 16.2: Info Adex metadata

Atribut	Tipus	Descripció	Exemple
Sector	String	Especifíca el sector en el que és troba el producte anunciat	MEDIOS DE COMUNICACION Y TE-LECOMUNICACIONES
Marca	String	Especifíca la marca principal de l'anunci	RASTREATOR

Taula 16.3: Kantar TNS metadata

Un cop presentades les definicions ens n'adonem que realment els camps plantejats si que expressen un mateix concepte i per tant, que són integrables. L'únic problema existent actual resulta ser moltes de les relacions no es mantenen i per tant, s'ha de prendre una decisió sobre si mantenir-les en la integració i donar-les com a suposició en els casos en els quals no existeixen o eliminar-les i perdre informació en els casos en què existeixen. Després d'una conversació ràpida amb l'Analista Programador i un dels Estadístics, s'ha decidit prescindir de la relació. Aquesta decisió s'ha pres perquè s'ha considerat que no redueix la qualitat de les dades més completes de manera molt significativa i ofereix més flexibilitat de cara a possibles conversions que es tractaran més endavant. A més a més, existia la possibilitat que el fet d'afegir dades a algunes de les fonts resultés en una possible alteració de les mateixes i com a conseqüència en una desviació de la realitat.

Així doncs, el diagrama UML de classes final de les fonts de dades globals al que s'ha arribat una vegada comprovats els camps que poden ser creuats és el mostrat en la figura 16.4.

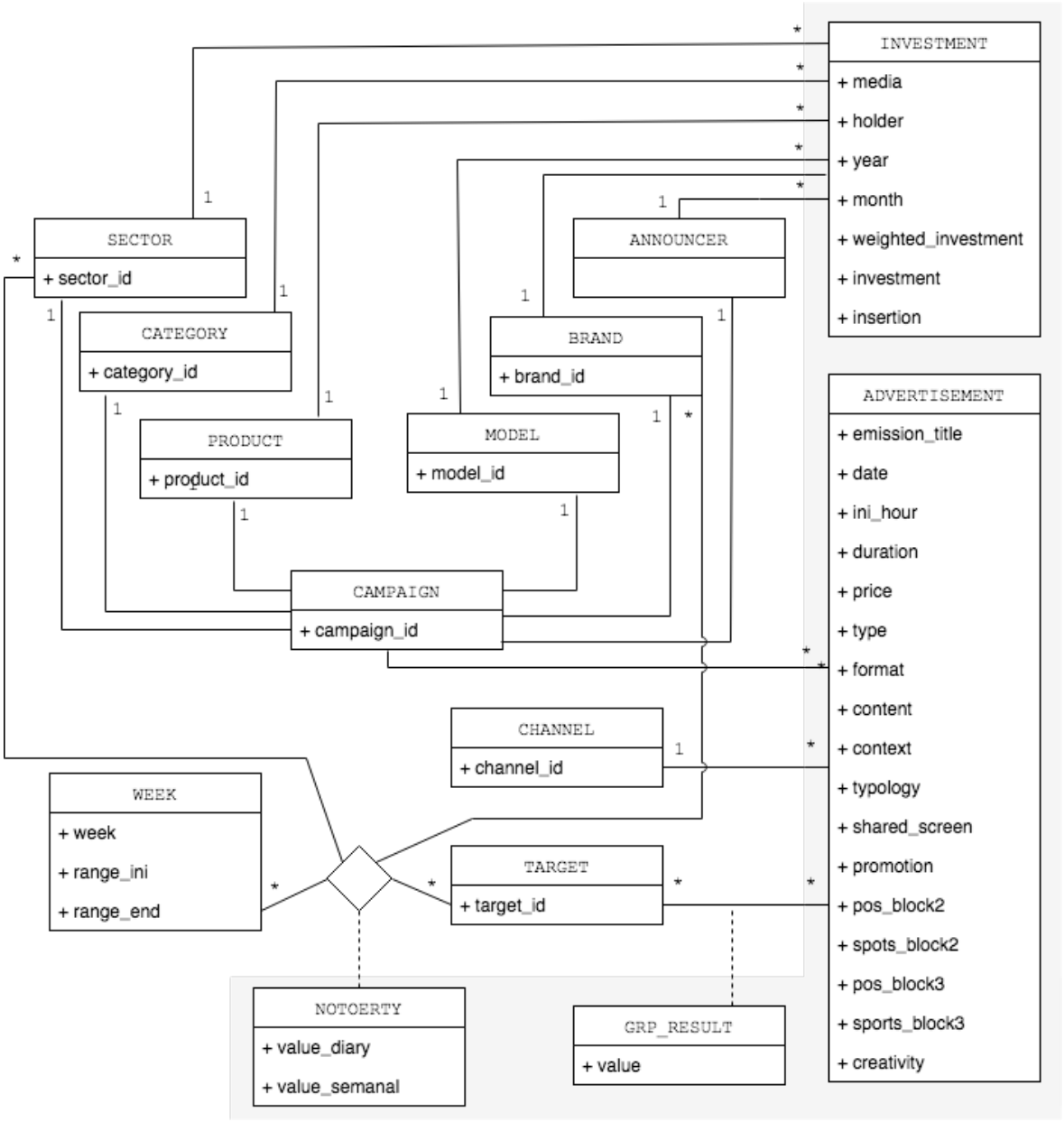


Figura 16.4: Diagrama de classes UML de les fonts globals integrades

Un cop exposat el diagrama UML de classes global integrat veiem que el que s'ha fet principalment ha estat agafar les classes que s'ha detectat que són les mateixes per a les fonts de dades i mantenir-les com una sola classe (és a dir les classes: sector, announcer, category, brand, product, model i target). A més a més, s'han eliminat les relacions d'aquestes classes, fet que ha fet incrementar el nombre de vincles que apareixen al diagrama, ja que, aquests vincles trencats s'han de passar a representar com a vincles directes cap a la classe que representaven.

Concloem doncs aquest subapartat confirmant que, s'ha trobat un diagrama de classes que integra totes les dades amb independència del valor (únicament tenint en compte la definició) i passem al següent pas per afegir la dependència del valor.

16.1.2 Problema valor vs. definició

Com ja s'ha pogut veure abans, la definició dels camps treballada a la taula de meta dades (punt d'estudi de les fonts de dades) encaixa. Però, podem veure que els valors de les dades, tot i representar un mateix objecte no és igual i per tant, l'encreuament no pot realitzar-se directament sinó que requereix una conversió prèvia.

S'ha trobat una solució a aquesta problemàtica basada en dos conceptes bàsics per al tractament de dades. Un son els estàndards i un altre les conversions. Pel que fa als estàndards de cara a l'encreuament de dades resulta ser un concepte bàsic, ja que si unes dades no estan estandarditzades i mantenen diferents formats i formes no podran ser mai encreuades, ja que, no existiran vincles a partir dels quals realitzar l'encreuament. Per tal de poder fer l'encreuament necessitem l'altre concepte bàsic que és la conversió, ja que, quan unes dades estan en un format no estandarditzat i volem estandarditzar-les l'acció a realitzar és convertir les dades per tal que segueixin el format desitjat.

A partir d'aquests dos conceptes s'ha generat la idea amb què solucionar aquest problema. Mitjançant 3 capes. Una primera que resulta ser el valor no estandarditzat, una segona que representa la conversió d'aquest valor a un valor estàndard i una tercera que representa el valor estàndard. Presentem una figura amb un diagrama de taules que busca explicar com es pretén implementar aquest concepte (Figura 16.5).

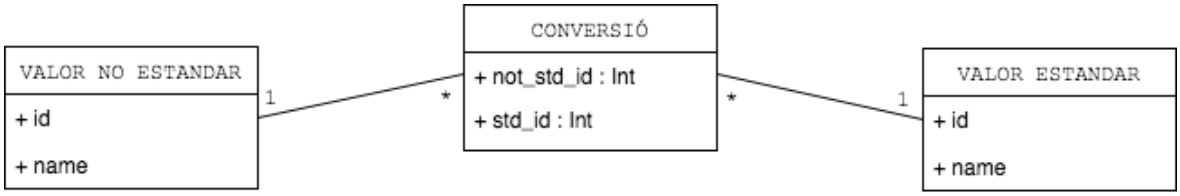


Figura 16.5: Solució al problema valor vs. definició

Veiem doncs que tot queda representat en 3 capes diferents. En la primera és guarda el valor que s'extreu de la font de dades, és a dir, un valor no estandarditzat. Existeix també una capa anomena davalor estàndard que mantindrà els valors estàndard i la capa intermèdia és la que es dedica a decidir per cada valor no estàndard quin valor estàndard se li assigna. D'aquesta manera aconseguim que un valor no estandarditzat pugui tenir un i només un valor estàndard assignat (fixem-nos que totes les relacions direcció esquerra-dreta resulten ser 1). Amb aquesta estructura aconseguim que, si cada font de dades conversiona les seves dades mitjançant els mateixos estàndard, una vegada acabada la conversió de totes les fonts, es pugui arribar a creuar mitjançant la capa estàndard de dades. Per tant, per a la utilització d'aquest mètode se substitueixen les antigues classes sector, announcer, category, brand, product i model per les classes sector estàndard, announcer estàndard, category estàndard, brand estàndard, product estàndard i model estàndard.

No s'expressarà aquesta solució a l'UML global integrat, ja que la quantitat de relacions que apareixerien seria elevada i no facilitaria la comprensió del concepte explicat. Aquesta proposta és veurà reflectida per primera vegada quan es realitzi la modelització de les taules de la base de dades. Concloem però, que s'ha trobat una solució al problema de què encaixin tant els valors com les definicions mitjançant la proposició plantejada.

16.1.3 Integració amb les dades extres de client

Una vegada integrades per complet les dades globals i havent trobat una solució al problema de valor vs. definició, procedim a realitzar una integració amb les dades del client. Per fer-ho ens basem amb la informació aportada pels stakeholders i per l'estudi de les fonts de dades extra. Veiem doncs que les fonts de dades dels diferents clients resulten no compartir format i que aporten dades diferents. Tot i aixo, de l'interacció amb els stakeholders i a força de preguntar mitjançant quin criteri o atribut creuaven aquestes dades actualment s'ha aconseguit detectar els dos casos més freqüents:

1. *Client que tan sols aporta dades pròpies:* Pel que fa als clients que tan sols aporten dades pròpies, a aquestes dades se'ls crea un nou atribut ja sigui: sector, announcer, category, brand, product o model. I a aquest atribut se'ls assigna el valor corresponent al client en les altres fonts. D'aquesta manera s'aconsegueix que al creuar mitjançant aquest atribut tan sols s'encreui amb les dades de l'empresa en si.

2. *Client que tan sols aporta dades múltiples:* Aquests clients solen aportar dades que contenen dades que contenen atributs amb una definició semblant a les definicions dels atributs sector, announcer, category, brand, product o model. A partir d'aquests valors es creuen manualment les dades amb les corresponents a les fonts globals de dades.

Per tant s'ha de trobar una solució que ens permeti incloure aquestes dades en l'estructura de classes ja generada en el cas de les fonts de dades globals. Recordem també que per a aquest cas no fa falta dissenyar un model UML de cada client sinó que tan sols s'ha de trobat mètode que permeti assegurar que totes les dades de client que compleixin una de les dues formes que s'han plantejat anteriorment puguin ser encreuades correctament.

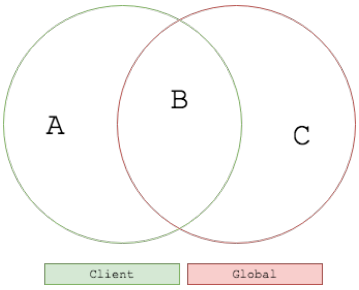


Figura 16.6: Possibles conjunts en l'encreuament de fonts globals i específiques

La solució que s'ha trobat és que per a tots els casos els valors estàndard siguin establerts a partir del client. És a dir, que les dades ofertes pel client siguin les que donen l'estàndard d'encreuament, aconseguint d'aquesta manera que les dades de client siguin creuables amb les altres tres fonts. Aquestes fonts globals convertiran els seus valors als valors donats pels clients. En cas que un valor no tingui conversió amb el client, és a dir, que es compleixi el cas C expressat a la figura 16.6, s'afegirà el valor a l'estàndard. Aquest fet no provocarà cap problema quant a l'encreuament amb el client, perquè com que no existeix aquesta dada en aquell conjunt, simplement es donarà un cas en el qual el encreuament amb les dades globals serà complet, ja que, tindrà les dades del client com a nul·les.

En el cas que es doni la possibilitat A, passarà el mateix però a la inversa, és a dir, que tindrem dades de client completes i les dades de les fonts globals amb valor nul. Finalment en el cas que totes les dades coincideixin, cas B, tindrem les dades encreuades completes tant de les fonts globals com de les fonts específiques de client.

Concloem doncs que amb aquest mètode de funcionament podem assegurar que les dades del client sempre seran creuables amb les dades globals recollides. I amb això ja podem assegurar que, tenim una estructura de disseny UML i unes descripcions de solucions als problemes d'aquesta que ens permeten tenir una estructura per a la integració de totes les dades.

16.1.4 Dependencia del client i de la font de dades

Una vegada donada per acabada aquesta descripció de l'estructura de classes necessària per encabir i representar les dades, es vol qüestionar un tema que ha estat molt repetit durant les reunions per part dels diferents stakeholders. Els principals stakeholders han estat molt reiteratius en comentar que tot havia de ser molt flexible i que cada client és un món i per tant, s'ha trobat de vital importància a aquesta alçada del projecte preguntar-nos si, donat un cas de conversió, aquest cas serà vàlid tan sols per un client o per més d'un client.

Com a resultat a aquesta pregunta s'ha rebut informació que indica que, aquesta suposició no es pot fer i que per tant, les conversions hauran de ser úniques, tant per client (1) com per font de dades (2). Per exemple:

1. Un client converteix producte 'Pelo rizado' a 'Rizos' i un altre no.
2. Una font de dades tracta sector 'Pelo' com a cabell i un altre com a depilació.

Per tant, caldrà que les conversions siguin font a font i que incloguin el client, ja que s'haurà de realitzar una taula de conversió per a cada font i client independentment.

16.2 Modelatge en forma de taules per a base de dades SQL

Una vegada definida una primera estructura de classes vàlida per al cas d'estudi i solucionats els principals problemes (restriccions textuals) passem a definir la solució en forma de taules relacionals que permetran encabir totes aquestes dades en una base de dades MSSQL. Per tal de fer-ho, com anteriorment s'ha fet amb la definició de l'estructura de classes, es seguiran un conjunt de passos explicats a continuació:

1. Traduir l'estructura de classes a estructura de base de dades amb les solucions comentades anteriorment.
2. Afegir taules que permetin gestionar l'historial de modificacions de les conversions.

Una vegada realitzats aquests passos es creu que es tindrà una estructura de taules vàlida per a poder mantenir totes les dades de les fonts globals i amb capacitat per a creuar aquestes dades i d'integrar-les amb qualsevol font de dades de client mitjançant el mètode prèviament explicat en el punt d'Implementació de les dades extres del client. Comencem doncs a traduir l'estructura de classes cap a estructura de taules per poder acabar obtenint la descripció de la base de dades necessària per a la solució proposada. Iniciem el procés de traducció de l'estructura de classes plantejada a estructura de classes. Per a aquesta traducció es tindrà en compte totes les restriccions i condicions textuals també explicades, és a dir, no

tan sols s'usarà el diagrama de classes UML plantejat a la figura 16.4. Resulta també important saber que MSSQL ofereix un sistema d'agrupació basat en 3 nivell: base de dades, esquema i taula. Aquesta estructura es farà servir amb aquest format [base de dades].[esquema].[taula] per poder facilitar el tractament i la comprensió de totes les taules. Així doncs l'exposició d'aquesta traducció es realitzarà per parts per a facilitar-ne la comprensió. Qualsevol taula que aparegui repetida, és a dir, dins d'una mateixa base de dades, esquema i amb el mateix nom, no existeix 2 cops en la base de dades sinó que s'ha hagut de mostrar per a demostrar una relació (com serà el cas de la taula [CBD].[std].[target]).

Per a facilitar una primera lectura superficial de les taules, afegim una descripció inicial dels diferents esquemes que es troben dins de la base de dades anomenada CENTRALIZED DATABASE [CDB]. Els esquemes a tenir en compte són diferenciables, ja que, estan identificats amb diferents colors. Trobarem els següents esquemes:

- *[std]:* Esquema en el qual s'emmagatzemen els valors estàndards de cada un dels atributs comuns entre fonts de dades. (Blanc)
- *[conv km]:* Esquema on es mantenen les conversions dels valors extrets de la font de dades KM. (Tronja)
- *[conv inv]:* Esquema que guarda les conversions dels valors de la font de dades INV. (Verd)
- *[conv km]:* Esquema que salva les conversions dels valors de la font de dades IOPE. (Groc)
- *[km]:* Esquema que manté la informació extreta de la font de dades KM. (Lila)
- *[inv]:* Esquema que guarda la informació de la font de dades INV. (Vermell)
- *km]:* Esquema que salva la informació obtinguda de la font de dades IOPE. (Blau)

A més a més a les taules apareixen els següents conceptes i abreviacions que s'explicaran a continuació per a facilitar-ne la comprensió:

- *PK:* Clau primària.
- *FK:* Clau forana.
- *UNIQUE:* Clau alternativa.
- *NOT NULL:* Valor nul no acceptat.
- *CHECK:* Comprovació.
- *MD5:* Algoritme d'encriptació.

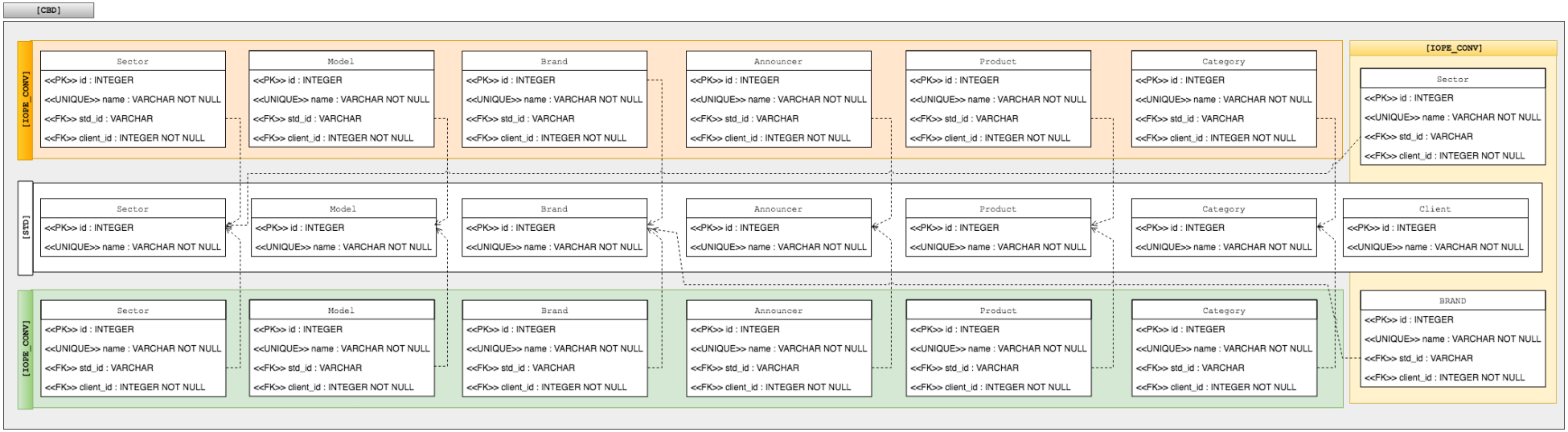


Figura 16.7: Traducció a UML de taules estàndard i taules de conversió.

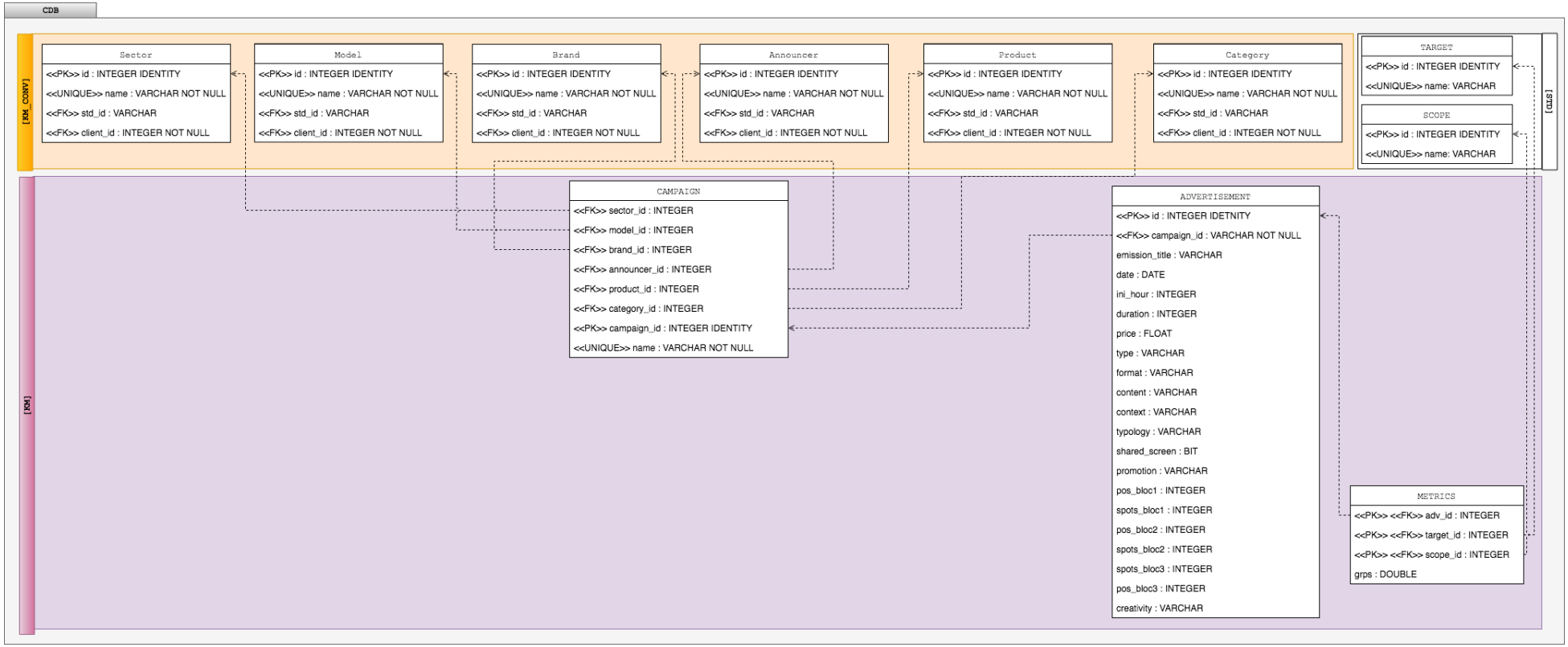


Figura 16.8: Traducció a uml de taules km i km conv.

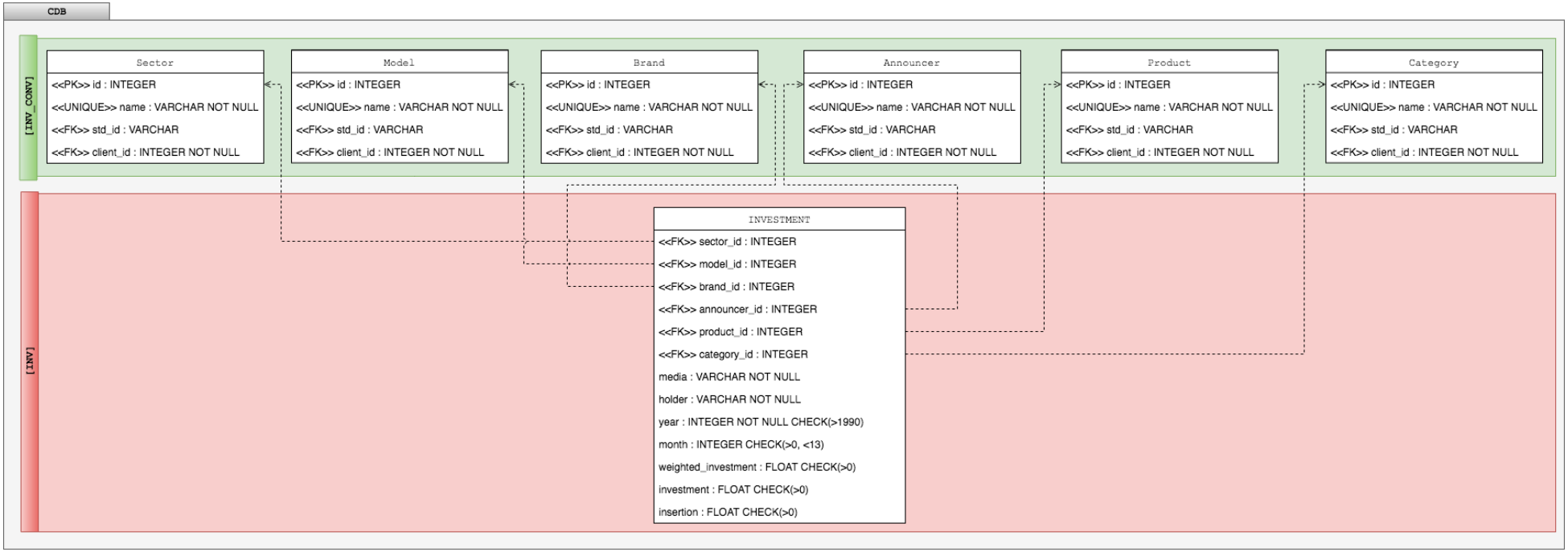


Figura 16.9: Traducció a uml de taules inv i inv conv.

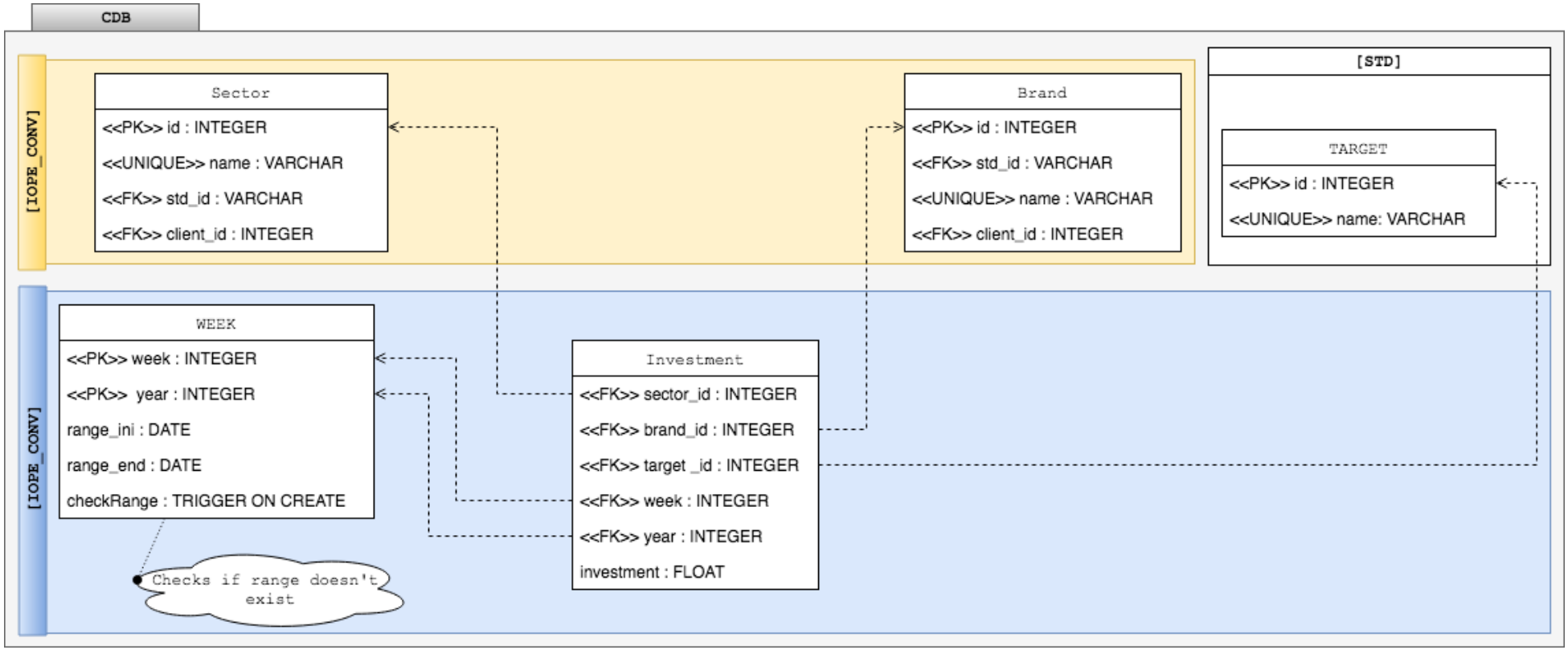


Figura 16.10: Traducció a uml de taules iope i iope conv.

Iniciem l'anàlisi de les taules presentades per les mostrades a la figura 16.7, que van directament relacionades amb el concepte explicat al punt de Valor vs. Definició. Amb aquestes taules podrem mantenir les conversions explicades al punt Problema valor vs. definició. Observem però que estem reduint el nombre de taules en 1 respecte de l'esquema ofert anteriorment gràcies a guardar el nom de la variable no estandarditzada a l'atribut nom de la taula de conversió que representarà el nom no estandarditzat de la variable, mentre que, la std id s'usarà per vincular-lo amb el seu valor estàndard. És a dir, std id serà la clau forana que apuntarà al valor id de la taula estàndard. Assegurem doncs amb això que, podrem mantenir les dades de les conversions de les taules globals per a l'encreuament i en cas d'omplir les taules d'estàndard amb els valors pertinents a la font de dades de client, aconseguirem que el encreuament sigui complet amb totes les fonts. Assegurant així que els requisits RF-F01, RF-F03 podran arribar a complir-se amb la implementació d'un front-end.

Ens n'adonem també a la figura 16.7 que, cada taula de conversió existent té el paràmetre client. Aquest fet és deu a què, com ja s'ha comentat anteriorment, les conversions no tenen per què ser exactament igual per a tots els clients, d'aquesta manera, forcem al fet que, s'hagi de generar una conversió per a cada client i assegurem que la proposta encaixi amb el cas d'estudi plantejat. Les relacions d'aquests atributs de clau forana amb la taula client representada a l'esquema [CDB].[std] no estan representades en el diagrama per a facilitar-ne la visualització, però, són existents.

Procedim ara a realitzar l'anàlisi de la figura 16.8 on podem trobar definit l'emmagatzamament de les dades extretes de la font de dades KM. Per a guardar la informació extreta d'aquesta font s'usa una taula anomenada campaign que mantindrà els diferents atributs relatius a una campanya vinculant-se via claus foranes a totes les taules de conversió existents a conv km. Posteriorment, aquesta taula es vincula amb la taula advertisment que mantindrà la informació relacionada amb els anuncis que s'han publicat durant la campanya, i a partir d'aquí i amb la relació amb target i scope (definint els conjunts de persones sobre els que es realitza l'anàlisi), apareix la taula metric que és l'encarregada de mantenir els valors de les diferents mètriques calculades.

Seguim ara comentat la figura 16.9 que descriu com ha de ser l'estructura de taules per aconseguir mantenir les dades extretes de la font INV. Veiem que aquest segueix realitzant la connexió de les dades amb les taules de conversió tot i que, en aquest cas, en comptes de connectar una taula externa a la informació, es relaciona directament amb la taula de dades, és a dir, la taula advertisement. Amb aquesta estructura doncs, podem assegurar que les dades de la font INV podran ser emmagatzemades.

Acabem parlant de la figura 1.6.10 que ens mostra les diferents taules que s'han de crear per tal de guardar les dades de IOPE. Per tal de fer-ho, s'hauran de tenir les taules week (representant una setmana a partir del nombre de setmana de l'any i l'any) que tot i poder semblar innecessàries (ja que el tipus DATE ofereix aquesta informació), resulten ser imprescindibles. Aquest fet és deu a què, els rangs de setmanes d'aquesta font no són estàndard sinó propis i canviants, així que molt probablement, s'haurà d'acabar treballant amb el rang de dades en comptes d'amb els valors dels atributs setmana i any.

Podem concloure doncs, que amb aquesta definició de taules, s'estan complint o ajudant a complir els requisits RNF-IE01, RNF-IE02, RNF-IE03, RNF-IE04, RF-F01, RF-FO2, RF-03, RNF-LD04. Aquest fet es basa en les explicacions realitzades anteriorment i més concretament, i a manera de resum, en què es poden mantenir les dades de totes les fonts i gestionar les conversions per tal d'acabar encreuant totes les dades.

Una vegada assegurat que podrem mantenir les dades extretes i encreuar-les gràcies al plantejament de taules realitzat, passem a afegir les taules que ens permetin mantenir un l'històric d'interaccions amb les conversions per tal de portar un control sobre els canvis que es realitzen sobre les dades.

16.2.1 Creació taules per a mantenir l'historial

Trobem de vital importància, que per tal de mantenir una gestió del sistema fàcil, resulta necessari identificar els canvis i tenir-ne un registre per a facilitar la marxa enrere d'aquests. A més a més, durant l'anàlisi de stakeholders i de procés realitzat ens n'hem adonat que, un dels problemes actuals és que, com que no existeixen tasques definides per a cada treballador, quan es troba un problema o un canvi en unes dades no es pot saber el motiu ni preguntar-ho a qui ho ha modificat, ja que, no existeix la possibilitat ni de comentar els canvis realitzats ni de localitzar la persona que ho ha modificat i preguntar-li. Així doncs i intentant encaixar amb els requisits RF-F03 i RF-F05, s'intenta trobar una manera que permeti:

1. Identificar als usuaris que realitzen una tasca.
2. Guardar el canvis realitzats.
3. Guardar justificacions dels canvis realitzats.

Per exemple, suposem que es realitza un canvi de conversió d'una marca que passa a tenir un nou nom estàndard, sigui UPC per exemple, que passa a dir-se Universitat Politècnica de Catalunya com a estàndard. Si el Joan (nom inventat per l'exemple) vol realitzar el canvi, aquest quedarà registrat en alguna part de la base de dades de manera que com a mínim, un altre usuari pugui saber que: "El Joan ha canviat el valor UPC per Universitat Politècnica de Catalunya a dia 25/07/2018, perquè així li ha demanat l'equip de direcció.". Per tant, s'hauran de generar una taula d'usuaris (per mantenir el qui) i com a mínim una taula de logs (per poder mantenir, que ha fet, que ha modificat i perquè). Així doncs, es proposa la següent estructura mostrada a la figura 16.11 de taules per a mantenir l'històric.

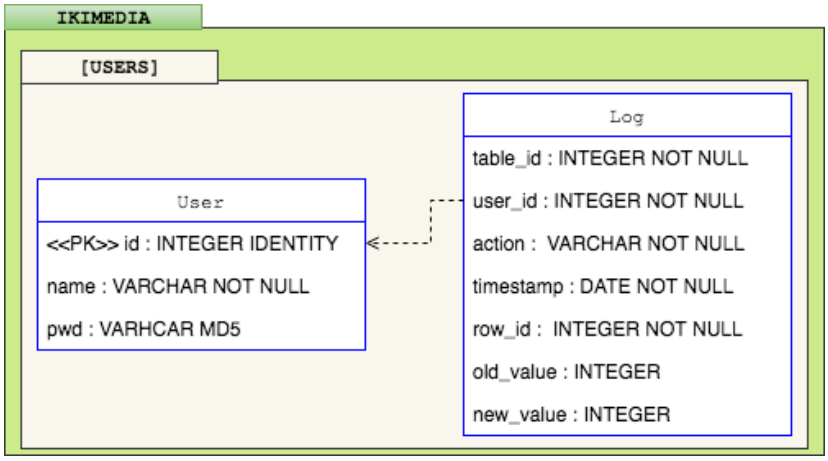


Figura 16.11: Proposta de taules per a l'històric i els usuaris

A la figura 16.11, podem veure que existeixen dues taules com prèviament ja s'havia suposat. Primerament, trobem una taula anomenada 'user' que busca donar la possibilitat de guardar els usuaris registrats en el sistema i que conté l'id assignat a l'usuari, el nom d'usuari i la seva contrasenya guardada amb una encriptació de tipus MD5 basada en claus privades per tal d'assegurar complir amb la llei de protecció de dades anunciada en el Boletín Oficial del Estado [**?**].

Pel que fa a la segona taula 'log', és una taula que pretén guardar l'històric d'interaccions amb el sistema de cada un dels usuaris. Per tal de fer-ho usa diferents claus foranes no referenciades què són:

- *table id:* Referència la taula a la que s'està realitzant un canvi. Aquesta referència pot ser basada en dues taules. La primera taula a la qual pot referenciar la taula [CDB].[sys].[table] què és una taula autogenerada per MSSQL o es pot generar una taula pròpia que emmagatzemi un registre de les taules actuals. En el nostre cas s'usarà la referència a la taula autogenerada per MSSQL.
- *user id:* Referència a la taula user, com podem veure a l'esquema.
- *row id:* Referència a la clau primària de la taula referenciada per table id. Aquest fet força al fet que totes les taules sobre les quals vulguem tenir un log necessàriament tinguin un id .autoincrementatiu. Fet que, encaixa amb les descripcions donades de les taules pertinents a [CDB].[std] i [CDB].[conv X].

Finalment, també podem veure que hi ha camps que permeten emmagatzemar la naturalesa del canvi, essent aquests: timestamp (dona la informació del moment de realització del canvi), action (descripció de l'acció realitzada), old value (valor anterior al canvi), nou valor (valor posterior al canvi).

Concloem amb aquest punt assegurant que amb aquestes especificacions, s'assegura la possibilitat de realització dels requisits RF-F03, RF-FO5 i RNF-AQ02 a nivell de back-end.

16.3 Lectures dels fitxers

Després d'una última reunió realitzada amb l'analista programador actualment s'ha decidit que totes les lectures no podran ser canviades i que forçosament hauran de ser realitzades a partir de fitxer .txt amb les dades extretes de la font que ell proporcionarà. Aquest fet que redueix les possibilitats d'automatització però queda justificat a causa de la gran quantitat de scripts que comenta tenir en execució que realitzen la lectura automàtica (sense tractament) i als que podria afegir una crida directa a la base de dades. Per tot això s'ha hagut de realitzar una modificació d'últim moment i en comptes de donar solucions basades en scripts temporals de lectura, simplement haurem de donar un subset d'aquesta solució i oferir procediments de la base de dades que siguin capaços de, donat un txt, llegir i actualitzar totes les taules de la base de dades. Per tal de fer-ho s'han plantejat dos casos, el cas de les dades globals i el cas de les dades específiques. Pel que fa a les dades globals s'ha generat els següent procés que pretenen realitzar la lectura i introducció de dades, reduint al màxim el temps de lectura i actualització de les taules de conversió amb l'objectiu de satisfer el requisit RNF-R01 i aturar el mínim de temps el sistema per a l'ús dels stakeholders.

Podem veure doncs a la figura 16.12 que aquest procediment es realitza en 4 passos i 2 processos. El primer procés està pensat per a realitzar-se el màxim de ràpid possible. A l'arribar al pas 4, ja ens permetrà treballar amb les dades. S'ha decidit que s'usarà una taula temporal en la qual s'emmagatzemarà amb una crida BULK tota la informació del .txt. El fet de realitzar la crida BULK en comptes de diverses crides INSERT no ens permet autoconvertir les variables d'encreuament a la id que les representa a la taula de conversió, ja que, realitzar la inserció en bloc no es poden tenir en compte les generated keys que un INSERT retorna com que, BULK no ofereix aquesta opció. Així doncs BULK es realitzarà sobre la taula temporal per aquest motiu, és a dir, perquè el format de les mateixes no podrà ser convertit directament. Un cop inserida tota la informació a la taula temporal, es passa a inserir a les taules de conversió tots els valors nous que es poden trobar en la taula temporal carregada anteriorment i una vegada finalitzat es procedeix a cridar l'altre procés.

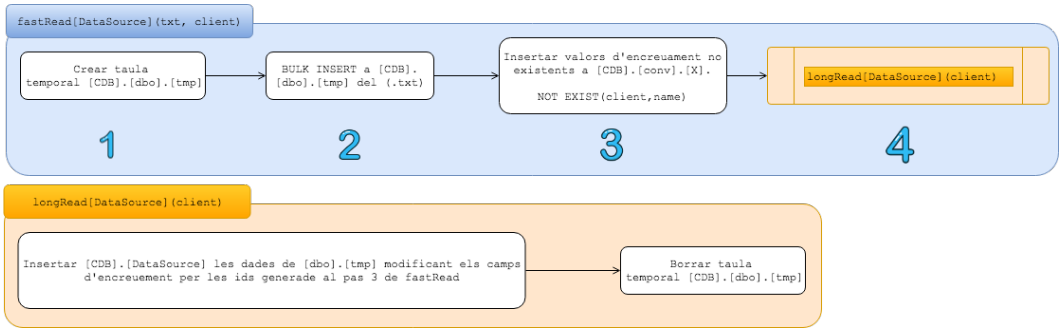


Figura 16.12: Proposta de procés de lectura de les fonts globals

El segon procés busca acabar de carregar les dades de la taula temporal a la principal. Per fer-ho canviarà els valors d'encreuament per les seves respectives id (PK i FK) i inserirà fila a fila. Per tant, perquè la inserció serà fila a fila i requerirà cert nombre de joins (depenent de la font), aquest procés resultarà més lent. Tot i això, al no impedir que es pugui treballar en les conversions no resultarà un alentiment rellevant de cara al producte final que es busca oferir. Un cop acabada aquesta lectura s'esborrarà la taula temporal donant així permís a poder realitzar una altra execució dels procediments. Recordem que en cas d'executar una lectura mentre se n'està realitzant un altre, en existir la taula temporal (no hauria estat esborrada) donaria error i no s'executaria.

Finalment, pel que fa als casos de les fonts específiques el procés és exactament el mateix però canviant l'esquema de conversió pels valors estàndard com s'ha plantejat al punt 'Integració amb les dades extra de client'. (Figura 16.6)

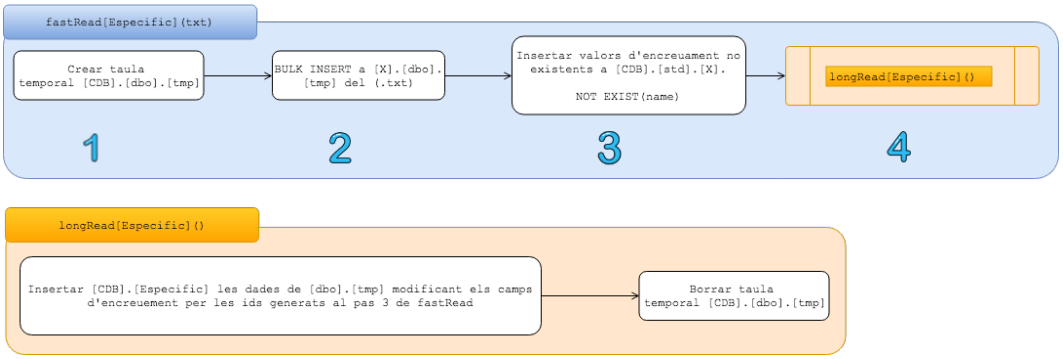


Figura 16.13: Proposta de procés de lectura de les fonts específiques

16.4 Disseny d'una interficie

Finalment i per acabar la proposta, s'ha d'incloure el disseny d'un front-end que demostri que es pot complir amb totes les funcionalitats demanades mitjançant l'eina de gestió. Per tal de fer-ho, es seguirà el següent patró, es presentaran les diferents vistes proposades i es demostraran les interaccions de cada una d'elles amb el back-end (mitjançant artefactes com els diagrames de flux). Per tant, si, un cop exposades totes les pantalles, es demostra que es poden executar tots els casos d'ús, podrem concloure que la interfície encaixa amb les necessitats reals que suposa que ha de solucionar.

Un altre punt a tenir en compte és que, per a la realització d'aquest front-end s'ha decidit usar una estructura MVP no bloquejant, és a dir, una estructura que no bloquegí el sistema al realitzar una acció sinó que l'executi en segon pla mitjançant processadors diferents dels de la interfície gràfica. Aquest fet és deu a què en el requisit RNF-R01 es va demanar que fos el mínim bloquejant possible i que permetés treballar sense haver d'esperar al fet que totes les tasques de lectura o de modificació es completessin. La figura següent mostra l'estructura de classes necessària per a la implementació d'aquest disseny.

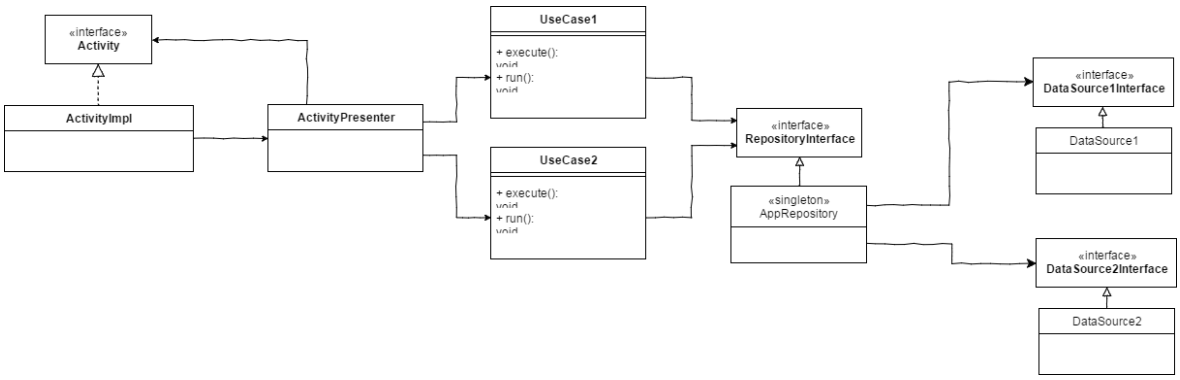


Figura 16.14: Disseny MVP

Continuem mostrant el flux complet de mock-ups tant de l'usuari bàsic com de l'administrador per tal d'oferir una visió genèrica del flux d'interacció que l'usuari haurà de fer per tal d'aconseguir tots els casos d'ús. També ens servirà per demostrar que tota l'aplicació segueix un mateix patró de disseny, fet que, resulta necessari doncs fa que la corba d'aprenentatge de l'usuari sigui més baixa i que per tant el programa sigui més usable. Fet que, resulta necessari com mostra el requisit RNF-R02.

Exposem doncs les següents dues figures (Figura 16.15 i 16.16):

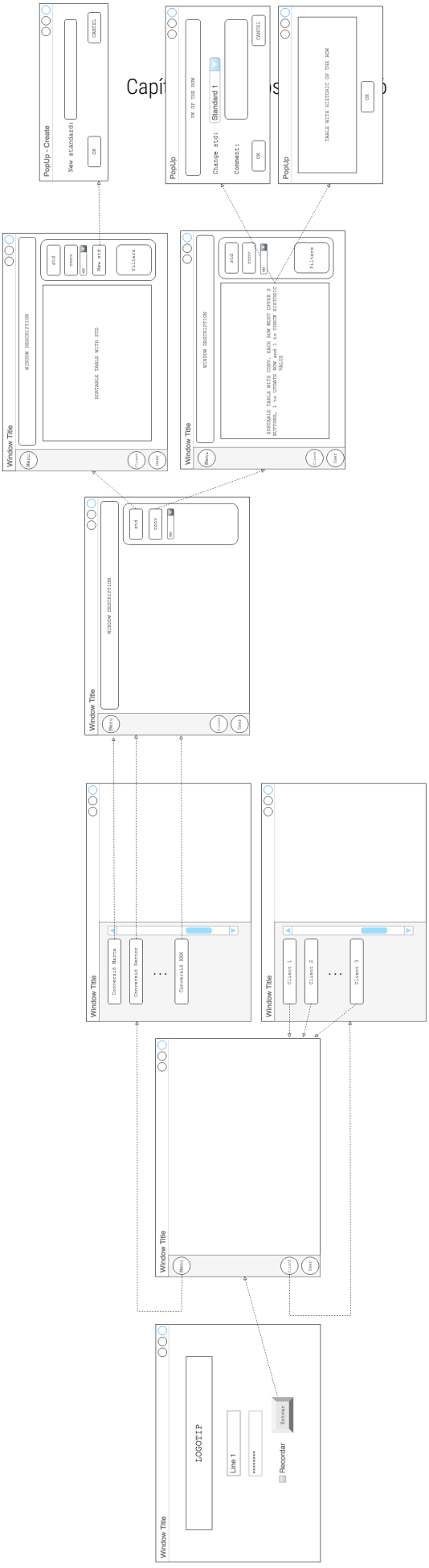


Figura 16.15: Mock-ups de l'usuari bàsic.

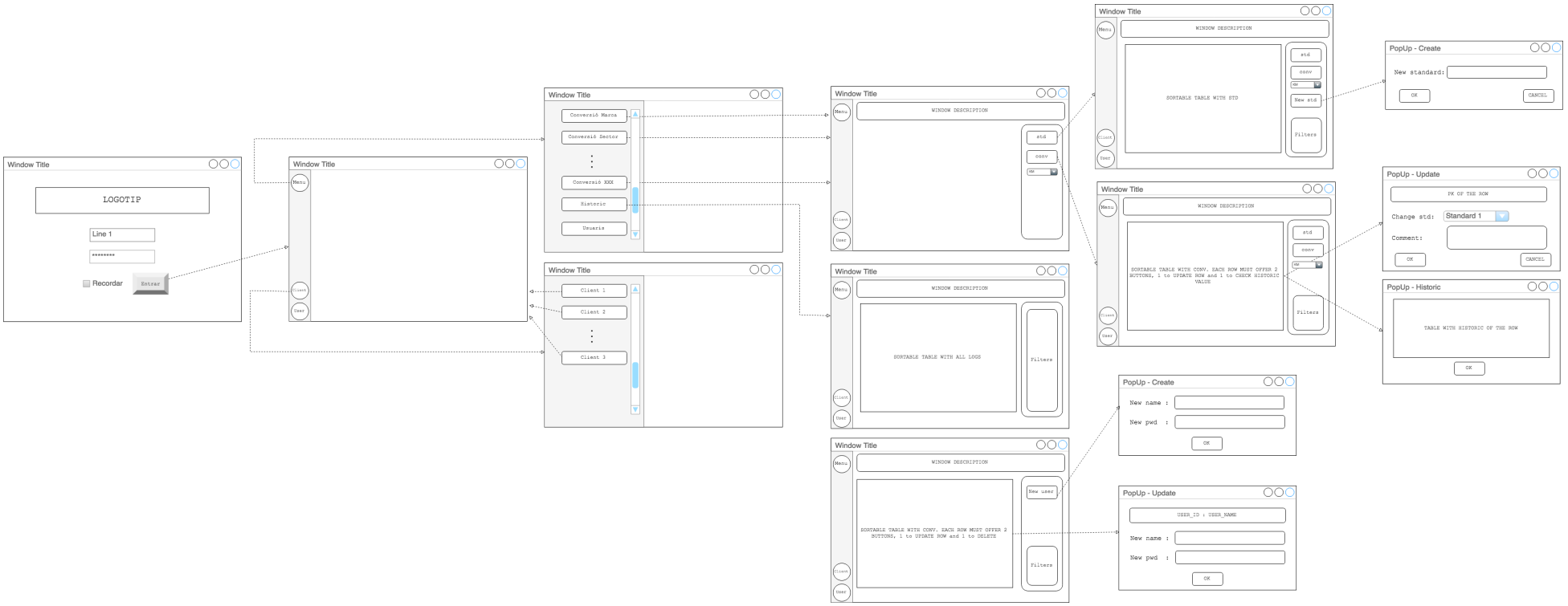


Figura 16.16: Mock-ups de l'administrador

Comprovem doncs que, com s'ha comentat abans en ambdós casos, és compleix un cert patró de disseny de totes les finestres permetent així que resulti més intuïtiu d'usar i més fàcil d'aprendre. A més a més, es pot veure també que ofereix sempre una barra lateral que informa l'usuari del client seleccionat i li recorda amb quin usuari està registrat en l'aplicació. La primera ens servirà per aconseguir que l'usuari tingui sempre present que la feina que està realitzant tan sols és vàlida per a un client, sigui quina sigui la pantalla. La segona, en canvi, servirà per recordar-li a l'usuari què s'està enregistrant les seves accions (no per atemorir-lo sinó per aconseguir que no realitzi canvis no justificats). A part d'això, en totes les finestres existeix un requadre superior que ens indica on ens trobem. Amb aquest requadre es busca disminuir la sensació d'estar perdut per les finestres d'una app de cara a millorar la usabilitat del programa.

Procedim ara a realitzar l'exposició pantalla per pantalla per a comentar, com prèviament s'ha exposat, la intenció de la pantalla i quins casos d'ús està resolent. Iniciem aquest procés amb l'exposició de la pantalla de log in a la figura 16.17.

Figura 16.17: Mock-up log in (W00)

Aquesta és una pantalla simple que permetrà l'autenticació dels usuaris mitjançant nom d'usuari i contrasenya. També existirà un checkbox anomenat Recordar que mantindrà aquest usuari i contrasenya guardats per evitar que l'usuari els hagi d'introduir cada vegada. Així doncs, el flux que es seguirà en prémer el botó d'entrar serà el següent:

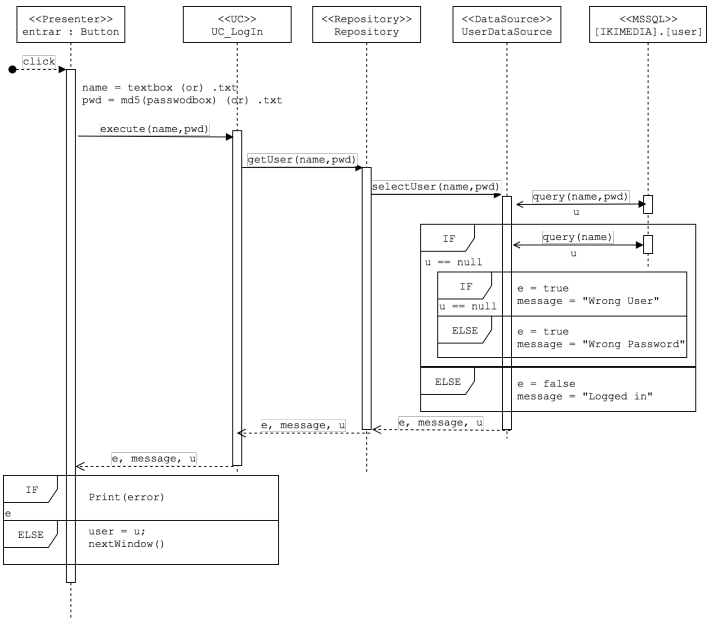


Figura 16.18: Procés de log in

Passem a exposar ara la pantalla d'inici a la que se'ns emportarà una vegada ens hàgem autenticat. Val la pena recordar que en aquest moment el programa ja recordarà el nostre usuari i que per tancar la sessió tan sols haurem de tancar el programa i tornar-lo a obrir per a poder iniciar sessió de nou amb normalitat.

Figura 16.19: Mock-up pantalla inicial [access-bar] (W01)

A figura 16.19 exposada ens trobem amb la barra lateral, l'anomenarem 'access bar'. Té 3 botons diferents que mostren imatges. El primer començant per baix mostra la imatge de l'usuari per saber qui està autenticat, passant el cursor per sobre d'aquest botó s'obre una pestanya amb la informació de l'usuari. El segon mostra la icona del client seleccionat, i, en cas de clicar sobre aquest botó, obre un menú desplegable que permet seleccionar-ne un altre. Per l'aparició d'aquest menú, es seguirà una navegació com l'exposada en la figura 16.19.

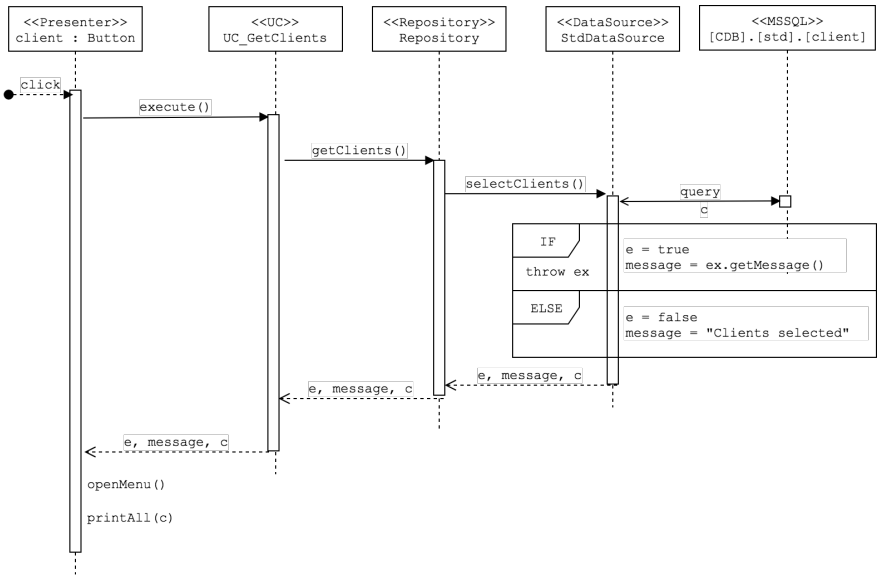


Figura 16.20: Procés d'obtenció dels clients

El menú desplegable que hem clicat, s'obrirà amb la funció openMenu() exposada al diagrama de navegació i quedarà la següent pantalla:

Figura 16.21: Mock-up menú del client (W02)

Aquest menú es carregarà amb els clients obtinguts de la query realitzada i servirà per seleccionar el client amb el qual treballar, en clicar un client diferent del seleccionat, es posarà la seva imatge en el botó de la 'access bar' i es recarregarà la pantalla des de l'inici (per evitar que estiguem en la taula de conversió amb un client seleccionat diferent de la informació mostrada, aquest concepte s'entendrà més endavant).

Finalment, l'últim botó que apareix simplement obra un nou menú desplegable (que anomenarem menú principal) que ens ofereix les diferents opcions de pantalles de treball. Aquest menú es veu identificat a la següent figura (cas del menú d'administrador, ja que engloba ambdós casos) (Figura 16.22):

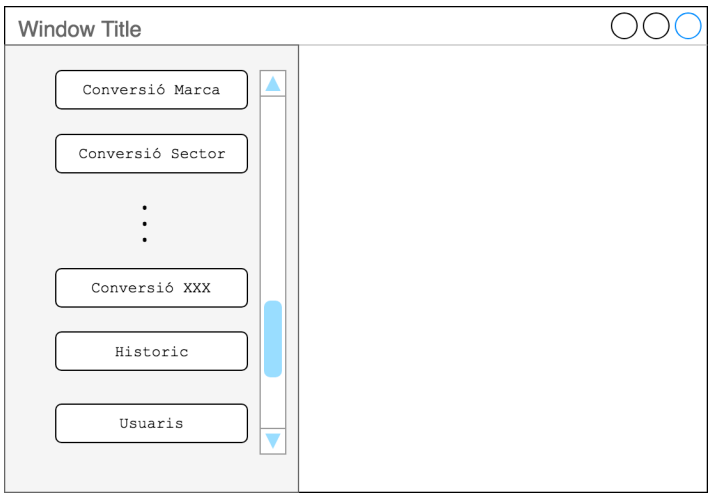


Figura 16.22: Mock-up menú principal [admin] (W03)

Tornem a veure en el menú principal presentat que existeixen 3 tipus de botons ('Conversió' comú a tots els tipus d'usuari i 'Històric' i 'Usuaris' que tan sols és d'administrador). Comencem mirant el cas de conversió. En el cas de clicar a conversió, es tancarà el menú i s'obrirà una nova finestra en la qual es podrà treballar amb el tipus de conversió seleccionat. És a dir, en cas de clicar conversió de sector, podrem treballar tan sols amb els sectors i el client seleccionat. La figura 16.23 mostra la finestra que ens apareixerà en cas de clicar a 'Conversió.'

Figura 16.23: Mock-up taules de conversió (W04)

Observem que aquesta finestra conta de tres elements. Un identificador de finestra en el que hi haurà mostrat la pantalla en la qual ens trobem, per exemple, Çonversió Marca", una combobox i dos botons. La combobox ens servirà per seleccionar la font global sobre la qual volem treballar, és a dir, sobre quina font de dades volem que se'ns mostrin les possibles conversions. Aquest fet no alterarà la mostra dels valors estàndard que són independents de les fonts globals. Pel que fa al primer botó, s'anomenarà std i ens obrirà una taula a l'espai en blanc de la finestra actual i botons relacionats amb la creació d'estàndards i filtres. A més, aquest botó modificarà l'identificador de finestra afegint "(std)"per recordar a l'usuari que està tractant amb estàndards. La taula que farà aparèixer aquest botó serà omplerta seguint el flux mostrat a la figura 16.24.

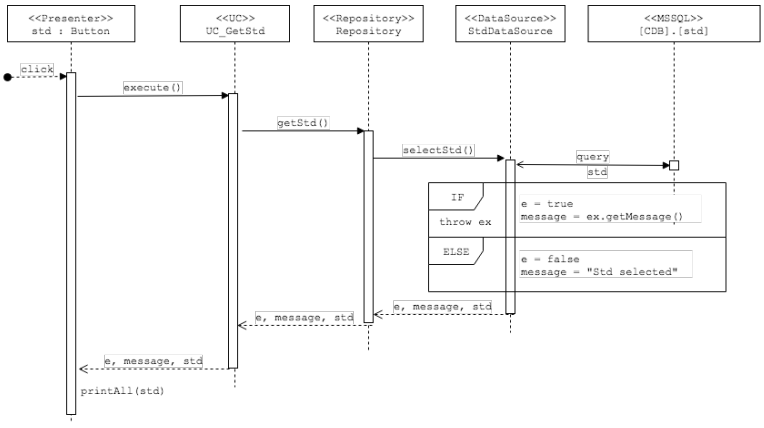


Figura 16.24: Procés d'obtenció de les conversions [Std]

Una vegada carregada aquesta informació i amb l'aparició dels diferents components explicats prèviament doncs quedarà una pantalla com la mostrada a la figura 16.25.

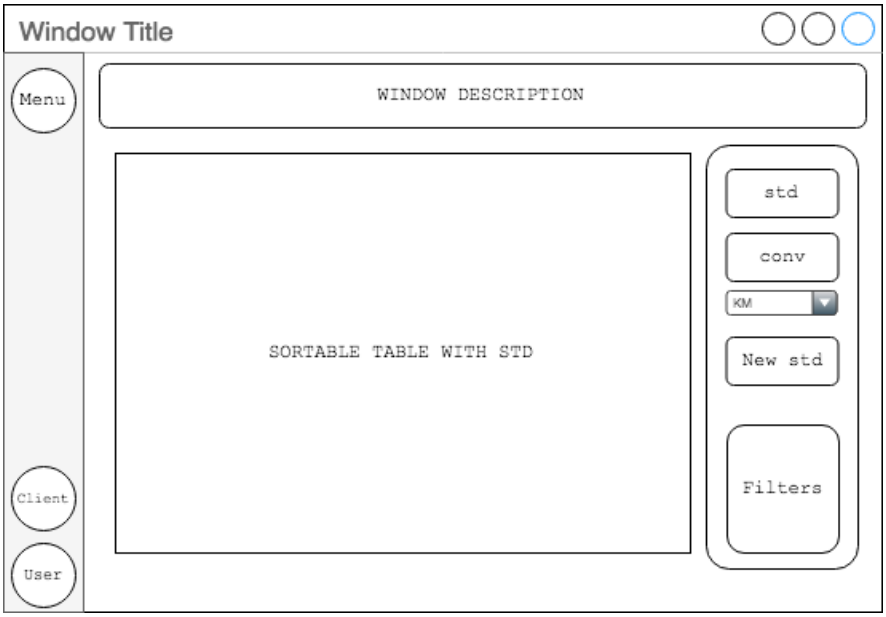


Figura 16.25: Mock-up de les taules de conversió [Std] (W05)

Com podem veure en aquesta pantalla tindrem opció de filtre que ens permetrà realitzar un conjunt de filtres sobre la taula per a millorar la usabilitat del programa i facilitar les tasques a l'usuari. A més a més, com s'ha comentat anteriorment també apareix un botó que ens permetrà crear nous valors estàndard. Per fer-ho tan sols s'haurà de fer clic i s'obrirà un pop-up amb les instruccions de creació d'un nou valor estàndard. Aquest pop up es veu mostrat en la següent figura (figura 16.26):

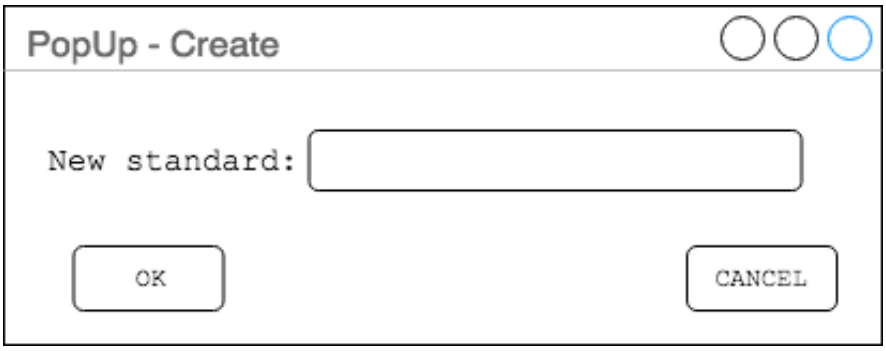


Figura 16.26: Mock-up pop-up de la creació d'estàndards (WO6)

El pop up de creació d'estàndard com podem veure a la figura 16.26 tan sols ens demanarà que introduïm el nom que volem donar a l'estàndard i dues opcions, cancel·lar i acceptar. En el cas de cancel·lar tan sols es tancarà el pop-up, en canvi en el cas d'apretar acceptar, es crearà un nou estàndard mitjançant el següent procediment (figura 16.27):

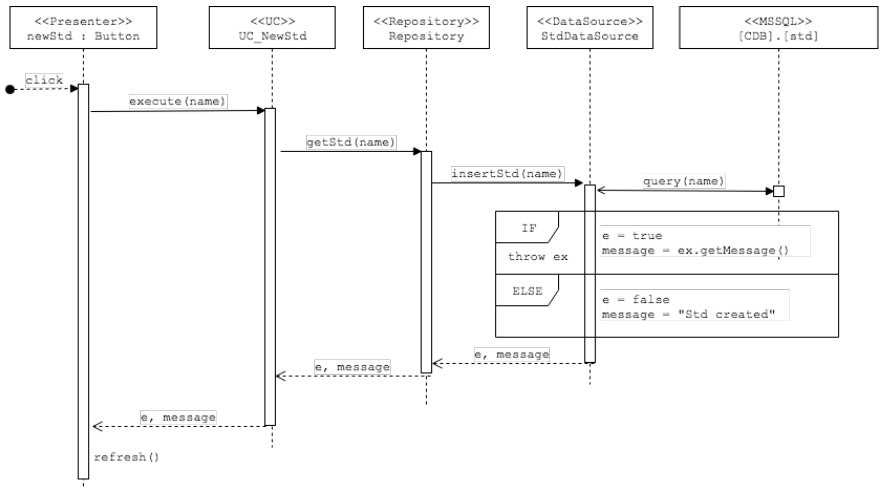


Figura 16.27: Procés de creació d'un estàndard

Tornem ara enrere fins a arribar altre cop a la pantalla inicial de conversió (W04) en la que podíem triar entre 'std' o 'conv' i seleccionem conv. Aquesta selecció dispararà un canvi de l'identificador de la finestra agregant al nom l'abreviació "(conv)ï entre claudàtors el nom del client seleccionat (recordem que, les conversions són per client). També s'executarà el següent procés (figura 16.32) per a carregar la taula corresponent. (Val la pena recordar que, aquesta càrrega serà feta sobre l'esquema çonv"+ valor de la combobox, és a dir, que tan sols carregarà les conversions de la taula relativa a la font de dades global seleccionada. Per evitar triplicar els processos de càrrega, modificació i consulta s'ha decidit realitzar una explicació general, ja que, tots funcionaran amb el mateix esquema però variant la font).

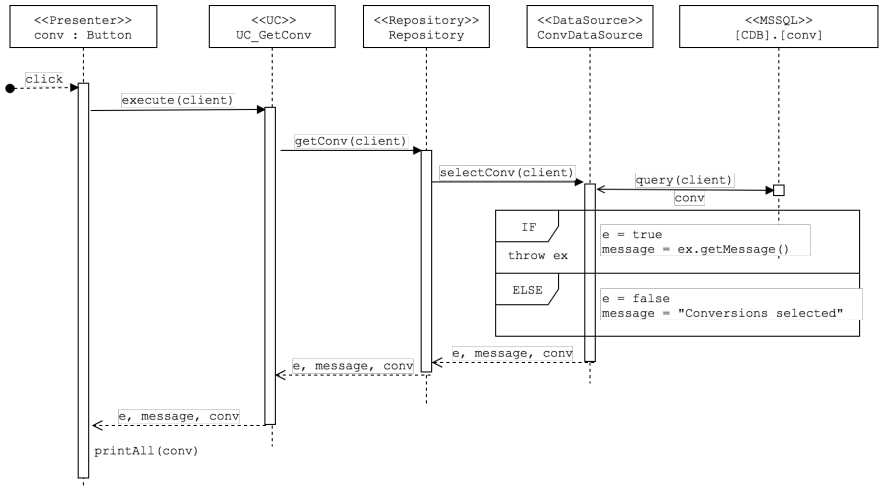


Figura 16.28: Procés d'obtenció de les conversions [Conv]

Al final d'aquest procés s'obtindrà una vista com la presentada a la figura 16.33 que permetrà a l'usuari interaccionar i gestionar les conversions. Resulta important recalcar que cada fila de la taula carregada tindrà dos botons, un de modificació i un de consulta que s'explicaran més endavant.

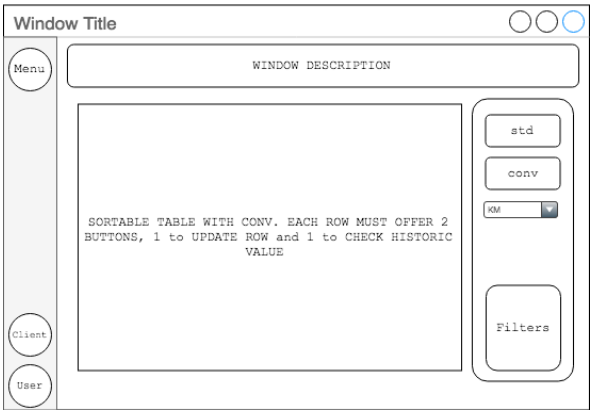


Figura 16.29: Mock-up de les taules de conversió [Conv] (W07)

Observem d'aquest mock-up que la taula és una taula ordenable i que conté les conversions relatives al client seleccionat. A més a més, cada fila ofereix els botons prèviament explicats que permetran interactuar amb aquestes. Comencem doncs parlant del botó de modificació que en seleccionar-lo obrirà el següent pop-up (Figura 16.30):

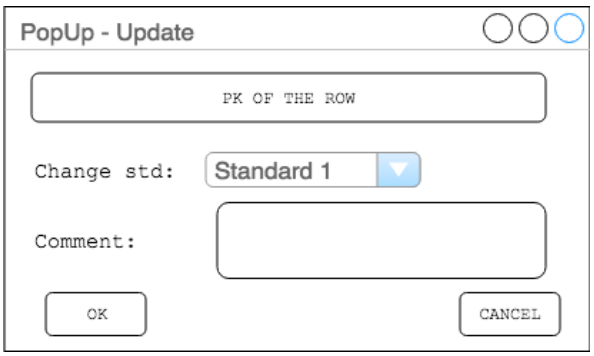


Figura 16.30: Mock-up pop-up de modificació de les conversions (W08)

En aquest pop up, trobem cinc elements. El primer resulta ser l'una referència a la fila que estem tractant (la clau primària/alternativa), el segon una llista de selecció mitjançant la qual es buscarà entre els estàndards existents per triar quin és el nou estàndard que es vol establir, el tercer un espai per a realitzar el comentari de justificació del canvi i finalment un quart i cinquè que són els botons d'acceptació o cancel·lació de la modificació. En cas de prémer el botó de cancel·lació, simplement es tancarà el pop-up, en canvi, si cliquem el botó d'acceptació, es dispararà el següent procés que, modificarà la conversió i guardarà a l'historial l'acció realitzada.

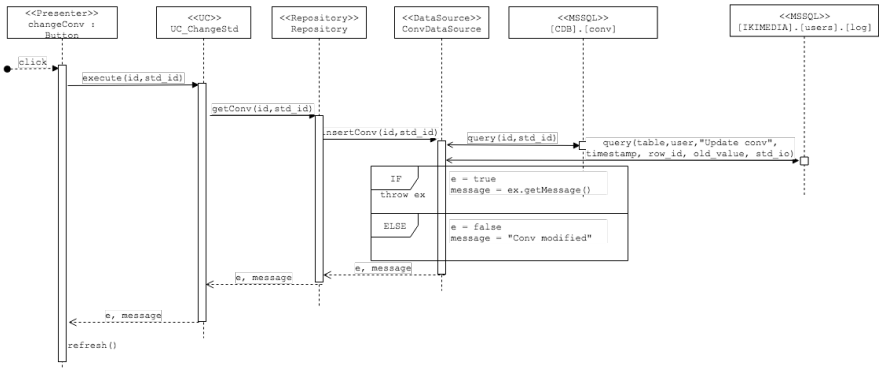


Figura 16.31: Procés de modificació d'una conversió

En canvi en cas de prémer el botó d'històric en comptes del botó de modificació, es disparà el següent procés (figura 16.32) que acabarà amb l'obertura d'un pop-up que mostrarà l'històric de la fila seleccionada amb l'objectiu de què l'usuari pugui consultar els valors que ha tingut una dada i entengui el perquè de certs canvis. Aquest fet pretén solucionar el problema d'assignació de responsabilitat de les tasques realitzades per més d'una persona.

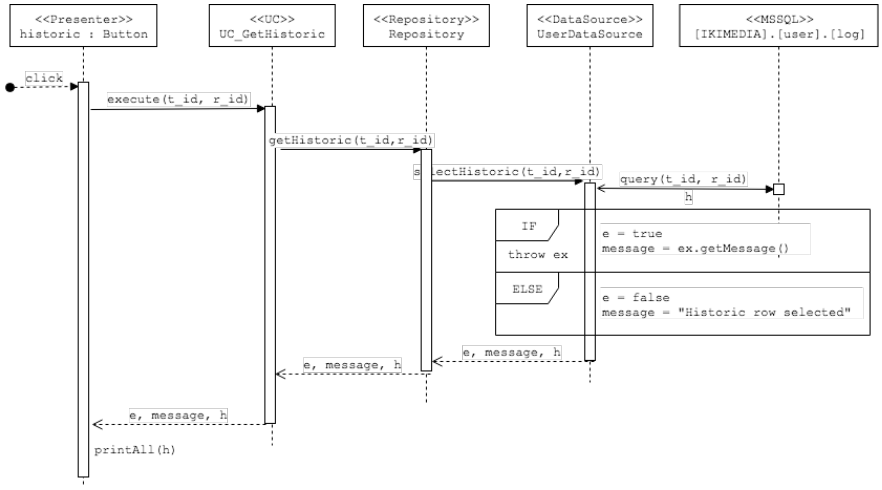


Figura 16.32: Procés d'obtenció de l'històric d'una fila

Un cop executat aquest procés (Figura 16.32) ens apareixerà aquest pop-up amb les dades recollides a la query per tal que ho consultem. Les dades apareixeran ordenades per data i la taula no serà ordenable ni filtrable. Podem veure la representació d'aquest pop-up a la Figura 16.33.

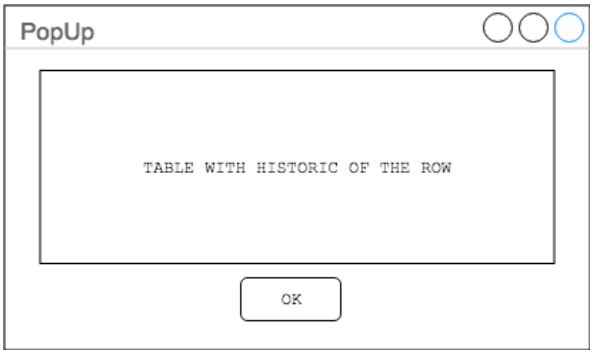


Figura 16.33: Mock-up pop-up d'obtenció de històric d'una conversió (W09)

Amb les vistes presentades fins ara acabem l'exposició de les opcions accessibles per l'usuari i comencem a afegir les vistes de l'administrador del sistema. Per exposar aquestes hem de tornar a la imatge oferta del menú principal (W05) i suposar que ara es selecciona el botó històric de la llista de botons. Aquest botó disparará el següent (Figura 16.34) procediment que busca obtenir la taula d'històric d'interaccions completa ordenada per data (timestamp).

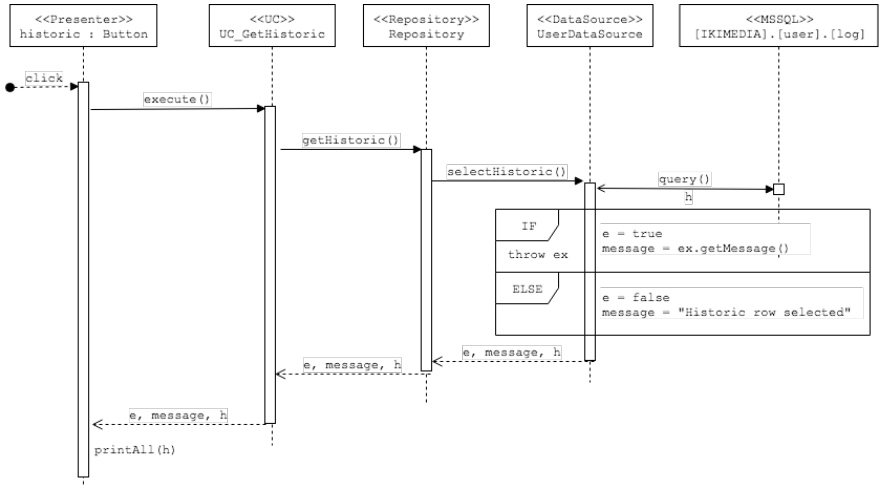


Figura 16.34: Procés de càrrega de l'històric complet

L'execució d'aquest procés acabarà amb el canvi de pantalla per passar a la pantalla que és mostra a continuació omplerta amb les dades obtingudes del procés. Podem veure aquesta pantalla mostrada en la Figura 16.35.

Com podem veure a la figura 16.35, aquesta pantalla tan sols disposarà d'una taula amb la informació obtinguda i una secció de filtres que serà prou completa per permetre navegar per la taula amb més facilitat. Aquest fet és degut al fet que, es busca que l'històric sigui immutable per tothom i que, aquesta eina tan sols sigui un mètode de consulta de cara a possibles incidències. Com més immutable sigui, més creïble serà.
Acabem ara passant a l'última opció a la qual té accés l'administrador des de el menú principal (W05), és a dir, a la gestió d'usuaris. En el moment en què es selecciona aquesta opció, es dispara el procés que és mostra a la Figura 16.36 que permetrà obtenir la llista completa dels usuaris registrats per l'administrador.

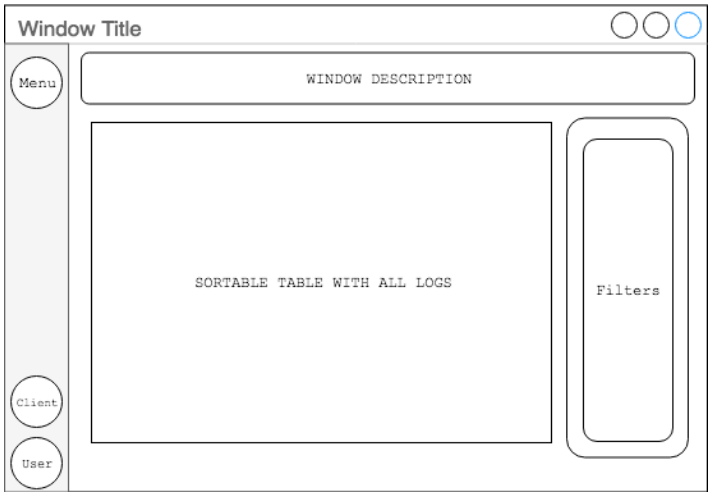


Figura 16.35: Mock-up de consulta de l'històric (W10)

Val la pena recordar que l'única persona amb capacitats de crear i administrar els usuaris serà l'administrador, aquest fet és degut al fet que, actualment en el sistema d'IKI Media Communications aquest és el funcionament establert.

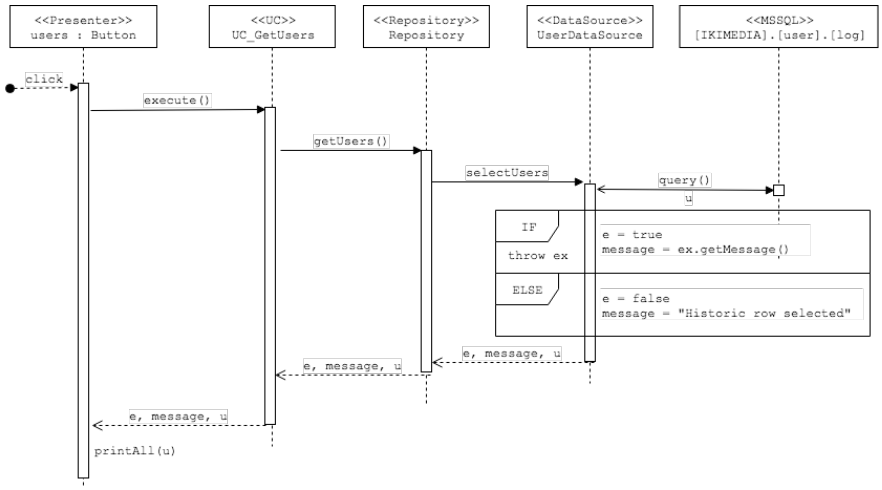


Figura 16.36: Procés d'obtenció dels usuaris

Un cop acabat aquest procés s'obrirà la finestra de la figura 16.37 mostrant totes les dades recollides, un botó per a la creació d'usuaris, un botó per a la modificació d'usuaris i finalment un seguit de filtres per a facilitar la cerca a la taula.

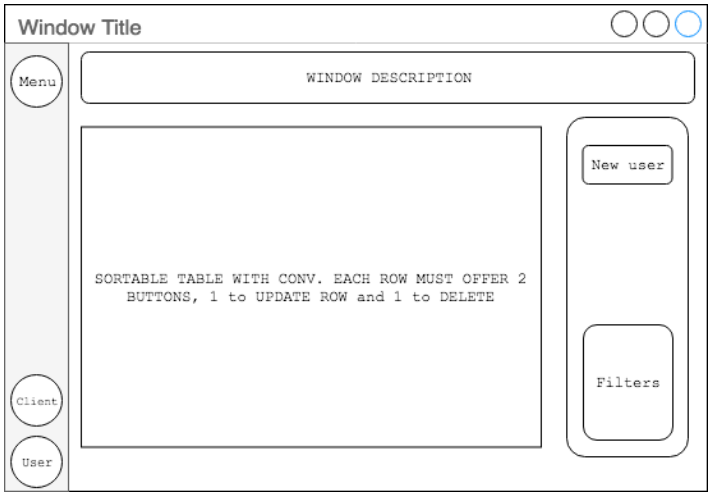


Figura 16.37: Mock-up de la gestió dels usuaris (W11)

Des d'aquesta finestra i en cas de clicar sobre el botó de modificació que trobarem a cada una de les files dels usuaris presents a la taula s'obrirà un pop-up com el mostrat a la figura 16.38 que ens permetrà modificar l'usuari.

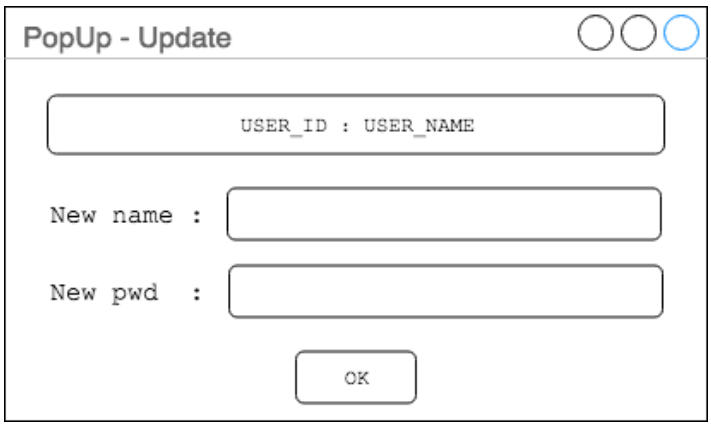


Figura 16.38: Mock-up pop-up de la modificació dels usuaris (W12)

Veiem en aquest pop-up que apareixen 4 elements. Un primer element on és mostra la clau primària de la fila seleccionada per recordar a l'usuari quin usuari està modificant i evitar errors, un segon i tercer que són textbox on podem introduir tant el nou nom d'usuari com la nova contrasenya (cal omplir els dos paràmetres per a modificar-ho) i finalment un botó d'execució del canvi que disparà el següent procés de modificació (figura 16.39):

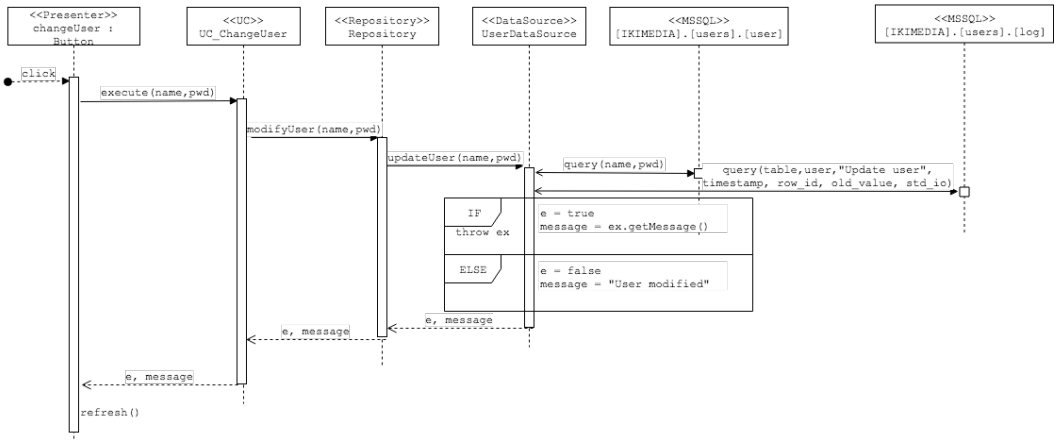


Figura 16.39: Procés de la modificació dels usuaris

Finalment acabem passant a presentar l'última opció del sistema què serà la de clicar des de la finestra inicial de la gestió d'usuaris (W12) a l'opció de creació d'un usuari, fet que obrirà el següent pop-up (Figura 16.40):

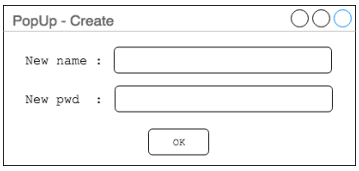


Figura 16.40: Mock-up pop-up de la creació dels usuaris (W13)

Podem veure que aquest mock-up resulta molt similar al mock-up del pop up (W13) mostrat, contenint tan sols una diferència que és que, en aquest cas no es mostra un espai de recordatori de la clau primària, ja que, resulta inexistent. Pel que fa a l'acceptar la creació, es dispararà el següent procediment que acabarà amb la creació de l'usuari final. (Figura 16.41)

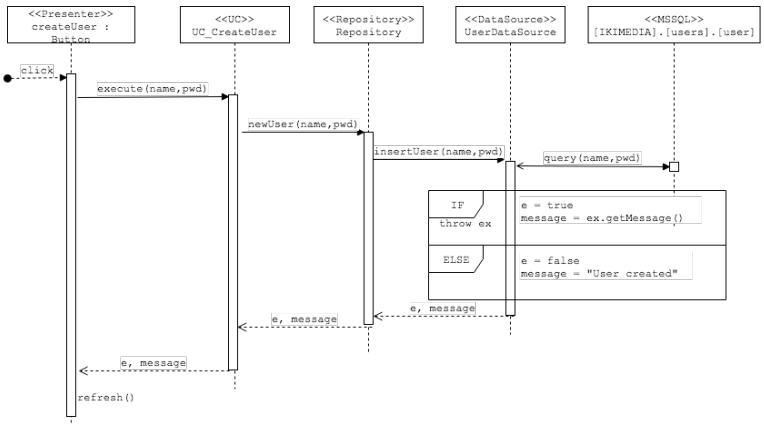


Figura 16.41: Procés de creació dels usuaris

Concloem doncs amb la presentació de les vistes i les respectives interaccions amb el sistema de back-end i passem a realitzar una taula de comprovació de si, aquestes vistes presentades encaixen amb els requisits funcionals establerts. D'aquesta manera, pretenem assegurar que, el treball fet, és de qualitat i que s'estan solucionant tots els casos d'ús existents. Per fer-ho ens basarem en el diagrama de casos d'ús, ja que, és una traducció directa dels requisits funcionals del sistema i resulta més entenedor. Tot i això també s'afegiran alguns requisits no funcionals que resultin rellevants.

Per realitzar aquesta anàlisi s'usarà una taula on apareixeran tots els casos d'ús (i els requisits no funcionals extres) i es justificarà en quines vistes s'estan complint. Comencem doncs la presentació de la taula (Taula 16.4 i 16.5):

ID Cas d'ús / Requisit	**Justificació**
UC00	• *W01:* Aquest cas d'ús es compleix a la pantalla de log-in. I pel que fa al tancament de sessió, és dóna per descomptat en totes les pantalles en tenir la possibilitat de tancar el programa.
UC01	Com s'ha dit anteriorment, a causa d'un canvi en els requeriments d'últim moment no serà necessari realitzar la introducció de dades via front-end, sinó que serà gestionada via procediments sql executables i per tant, no és un cas d'ús a complir en el front-end.
UC02	• *W09:* Pel que fa a la vista W11 ens permet consultar l'històric d'una fila de qualsevol taula de conversió mitjançant la clau primària. • *W10:* La vista W12 permet consultar tot l'històric de les accions rellevants realitzades sobre la base de dades completa.
UC03	• *W11:* La vista W13 permet consultar i eliminar tots els usuaris existents en el sistema. • *W12:* La finestra W14 ens dóna l'opció de modificar els usuaris. • *W13:* La pantalla W15 ens dóna permetrà de crear nous usuaris en el sistema.

Taula 16.4: Comprovació casos d'ús vs. Interfície. (Part 1)

ID Cas d'ús / Requisit	Justificació
UC04	• *W02:* El menú ofert per la vista W02 ens permetrà seleccionar el client sobre el qual volem treballar les conversions. • *W05:* La pantalla W05 ens dóna accés a les taules de valors estàndard. • *W06:* En aquest pop-up se'ns dóna la possibilitat de crear nous valors estàndard per a la conversió. • *W07* :La vista W07 ens ofereix una taula de consulta per als valors de les conversions d'un client determinat. • *W08* :Pel que fa al pop-up de la vista W08, ens permetrà modificar les conversions existents.
UC05	Pel que fa a aquest cas d'ús s'assegura pel fet que es pot complir el UC04 i per tant, amb l'existencia de les taules de conversió corresponents, es confirma que, l'usuari serà capaç d'aconseguir totes les dades encreuades mitjançant una sentència SQL basada en JOINS dels camps que s'han convertit.
RNF-LD02	El sistema d'interfícies proposat, pot ser implementat i està pensat per Java, més concretament, JavaFX.
RNF-LD05	Totes les aplicacions de l'empresa comparteixen un format similar estètic. Aquest fet també afavorirà la reducció en la corba d'aprenentatge.
RNF-P01	El sistema serà no bloquejant gràcies a la utilització de l'estructura de classes MVP que ens permetrà executar processos en segon pla, mentre el processador de la interfície segueix actiu.
RNF-P02	Com es pot veure en l'arbre de flux de pantalles presentat a les Figures 16.15 i Figures 16.16, el funcionament del sistema és bastant senzill. A més a més, s'aporten detalls com: les imatges de client (que aporten informació sobre quin client ens trobem treballant), referències clares sobre quin punt exacte del software ens trobem (per evitar la sensació d'estar perdut), etc.

Taula 16.5: Comprovació casos d'ús vs. Interfície. (Part 2)

Veiem doncs que tots els requisits funcionals queden satisfets i que els no funcionals relacionats amb aquest apartat de la interfície també queden tancats. Així que donem per vàlida la solució d'interfície proposada.

16.5 Conclusions

Una vegada acabades d'exposar totes les parts de la solució proposada, procedim a reafirmar la seva validesa completa de la proposta responent les tres preguntes exposades durant la introducció d'aquest mateix capítol. Trobarem doncs cada una de les preguntes respostes en una secció diferent de les presentades a continuació.

16.5.1 S'han usat els conceptes estudiats durant l'estudi de context?

Es pot comprovar com s'ha tingut en compte cada una de les parts de l'estudi realitzat, per a verificar-ho separarem cada una de les diferents tasques realitzades i justificarem on s'han usat per a comprovar que s'ha tingut en compte el context a l'hora de plantejar una solució, assegurant així que, la solució proposada encaixa amb el cas concret en el qual s'ha d'aplicar. Comencem doncs presentat la següent estructura amb les diferents tasques relacionades amb estudi de context.

1. *Estudi de les fonts de dades:* Aquest punt ha estat un dels temes més tractats quant al disseny d'un digrama de classes inclusiu i la traducció d'aquests. Recordem que s'ha partit de les meta dades i l'esquema de dades realitzat durant l'estudi de context per arribar a integrar-ho i aconseguir un esquema global que representes totes les dades. A més a més, gràcies a aquest estudi, s'ha pogut arribar a trobar possibles problemes en la integració, per exemple el punt valor vs. definició, i s'han aconseguit donar solucions que els ressolessin podent arribar a afirmar que totes les dades són encreuables i per tant integrables en un sol esquema de classes i per extensió, a un model relacional. Podem doncs comprovar que l'estudi de les fonts de dades ha estat usat i especialment rellevant per a la proposta de solució.

2. *Estudi del hardware:* Pel que fa a l'estudi del hardware, de cara a la proposta de solució, no ha estat especialment tingut en compte, doncs al tractar-se d'un sistema de base de dades MSSQL, com s'ha comentat en el punt d'anàlisi dels hardware i en el de selecció d'aquest, podem assegurar que podrà ser integrable en el servidor 'sql' existent. Per tant, concloem que aquest punt, tot i no estar tractat implícitament en aquest punt ha estat tractat explicitament, ja que, la solució presentada encaixa amb la descripció del hardware i no necessita modificar-lo per tal de funcionar des de qualsevol entorn. A més, la solució presentada, permet (al funcionar amb Java, .jar) ser usada des de tots els sistemes operatius existents a l'empresa (Windows i MacOs).

3. *Estudi del software:* De cara a l'estudi del software, ens trobem en una situació semblant a la de l'estudi del hardware. En aquest punt doncs, veiem que hi ha parts que no han estat tractades explícitament perquè ja han estat treballades en el punt de selecció del software. Tot i això, existeixen referències tant al llenguatge de programació (Java) com a el framework de desenvolupament d'interfícies (JavaFX), és a dir, tot el punt de disseny d'una interfície es basa en els llenguatges escollits per a la realització del projecte i han estat basats en funcionalitats que aquests ofereixen. A més a més, s'ha presentat una estructura de classes molt usada en Java per a la implementació d'Apps (MVP).

4. *Anàlisi dels stakeholders:* Podem afirmar que aquest punt ha estat tractat en totes les fases d'aquesta proposta, perquè principalment s'ha intentat adaptar les diferents visions dels stakeholders. Un exemple pot ser en l'apartat de creació d'un històric, que s'ha basat bàsicament en l'estudi de stakeholders realitzat, o durant el disseny UML de les fonts de dades específiques, on s'ha buscat una solució flexible que encaixes per a tots els clients. També podem veure en l'apartat de definició d'una interfície com s'ha buscat donar una interfície amb un flux fàcil per a reduir la corba d'aprenentatge, probablement amb la contraposicó de reduir la velocitat d'execució de les tasques. Això

s'ha fet per tal de facilitar als usuaris l'ús del software, ja que, pel vist durant l'estudi de context no resulten familiaritzats amb el món de la programació ni són usuaris freqüents de sistemes de bases de dades.

5. *Anàlisi de procés:* Aquest punt tan sols ha estat tingut en compte de cara a la generació de l'històric. S'ha trobat justificat realitzar aquesta part de disseny (disseny del 'log') perquè, al no tenir tasques definides per a cada treballador ha resultat d'una importància elevada generar aquest sistema de control intentant així reduir el nombre d'errors. Tot i això, aquest apartat serà tractat amb més profunditat en el següent punt de proposta d'un flux de treball.

Podem concloure doncs després de la justificació de cada una de les tasques que, l'estudi de context ha resultat útil de cara a la proposta de solució i per tant, que s'ha tingut en compte, assegurant així que la proposta encaixarà amb el cas real d'estudi.

16.5.2 Es compleixen tots els requisits?

Durant el transcurs de la proposta s'ha anat citant els requisits que s'anaven assolint amb cada una de les parts i per tant, podem afirmar que es compleixen. Tot i això, per a facilitar-ne la comprovació, oferim la següent taula amb la justificació de compliment de cada un dels requisits.

ID Requisit	Justificació
RNF-IE01	S'assegura que aquest requisit està complert, ja que, existeix un plantejament d'esquema relacional en el qual les dades poden encabir-se i perquè s'ha generat un plantejament de procés de lectura que assegura que les dades globals podran ser inserides en el sistema. Cal recalcar que, de cara al final d'aquest projecte s'ha realitzat un canvi en aquest requisit (per petició d'un dels stakeholders) que en reduïa l'abast a tan sols lectures de .txt.
RNF-IE02	S'assegura que aquest requisit està complert, ja que, existeix un plantejament d'esquema relacional en el qual les dades poden encabir-se i perquè s'ha generat un plantejament de procés de lectura que assegura que les dades globals podran ser inserides en el sistema.
RNF-IE03	Podem afirmar que s'ha donat solució a les lectures de les dades de client en el plantejament de la solució a la integració de les dades de client i, ja que, s'ha donat una proposició de procés SQL que permetria la lectura de les dades específiques.
RNF-IE04	Podem assegurar que totes les fonts de dades podran ser creuades donat que s'ha trobat un esquema de classes que integra totes les dades mitjançant les diferents taules de conversió i estàndar.

Taula 16.6: Justificació requisits: RNF-IE

ID Requisit	Justificació
RF-F01	Queda assegurat amb el procés de 'fastRead' que l'usuari tindrà sempre actualitzats els nous valors de conversió a omplir, ja que, aquest procés s'encarregarà d'introduir els casos de nova aparició. A més, el valor std id d'aquest serà nul i per tant, si l'usuari filtra per valor nul serà notificat de les noves aparicions al sistema.
RF-F02	El sistema dóna la possibilitat d'actualitzar una conversió, deixant rastre a l'historial, mitjançant la vista W08.
RF-F03	L'usuari serà capaç de comprovar els valors de conversió i l'històric mitjançant les vistes W07 i W09.
RF-F04	Es generarà un històric que ens permetrà controlar els canvis de conversió que es realitzin mitjançant l'execució de la vista W08.

Taula 16.7: Justificació requisits: RF-F

ID Requisit	Justificació
RNF-P01	Mitjançant l'ús de l'estructura de classes MVP assegurem que la interfície serà no bloquejant. A part, el procés de lectura de les dades està dividit en dues parts per a reduir el temps d'espera de la càrrega d'espera de cara a poder treballar sobre les dades sense tenir-les completament llegides.
RNF-P02	Durant tot el projecte s'ha buscat que la interacció amb el software fos el més simple i generes una corba d'aprenentatge el més baixa possible. Podem veure que el flux de navegació del software proposat és simple i segueix sempre un mateix patró facilitant així la interacció de l'usuari, a part, dels identificadors que trobem a cada pantalla o la simplicitat dels pop-ups presentats.

Taula 16.8: Justificació requisits: RNF-P

ID Requisit	Justificació
RNF-LD01	A l'estar usant una font de dades SQL (MSSQL), i com s'ha comprovat tant a l'estudi de context com a la selecció de dades, podem assegurar que serà integrable en R, Excel i Tableau.
RNF-LD02	El sistema dissenyat està plantejat per ser implementat en Java i JavaFX. A més s'usen estructures típiques d'aquest llenguatge de programació (MVP).
RNF-LD03	La implementació i posada en marxa d'aquest projecte no suposarà cap modificació en el hardware físic actual ni en els servidors virtuals existents.
RNF-LD04	La traducció de l'estructura és a model de taules relacionals i per tant, el sistema de bases de dades està pensat per a ser usat en SQL.
RNF-LD05	El conjunt d'interfícies segueixen el patró de disseny de l'empresa i un flux similar. Existeixen decisions basades en els patrons d'actuació ja establerts. (Aquest fet també redueix la curva d'aprenentatge)

Taula 16.9: Justificació requisits: RNF-LD

ID Requisit	Justificació
RNF-AQ01	El sistema resulta fàcil de mantenir un cop en funcionament, tot i que per condició aquesta proposta serà de desenvolupament continu (dades específiques).
RNF-AQ02	El sistema dóna una seguretat d'encriptació MD5 a les dades privades de l'usuari (en aquest cas tan sols la contrasenya).

Taula 16.10: Justificació requisits: RNF-AQ

Podem concloure doncs, després d'aquesta comprovació cas a cas, que tot els requisits plantejats es compleixen exceptuant el RNF-IE01, que ha estat modificat durant aquesta última iteració i se n'ha reduït l'abast.

16.5.3 Es pot implementar amb els softwares i hardwares usats?

Com ja s'ha comentat en les dues preguntes anteriors. Podem assegurar que, la proposta de solució donada pot ser integrada als softwares seleccionats i que no necessitarà cap tipus de hardware físic o virtual extra per tal de posar-se en funcionament.

Concloem doncs, una vegada respostes totes les preguntes plantejades que la solució proposada compleix amb tots els requisits existents, que encaixa amb el cas d'estudi (gràcies al fet que encaixa amb l'estudi de context realitzat), que pot integrar-se amb el software seleccionat i finalment que és implementable amb el hardware existent. Per tant, podem afirmar que resulta una proposta de solució vàlida.

Donem doncs per finalitzat aquest apartat i passem a demostrar quin hauria de ser l'ús d'aquest sistema per a poder realitzar les tasques fins ara portades a terme, millorant tant l'eficàcia com reduint el nombre d'errors per factor humà (recordem que el sistema proposat resulta ser en gran part automàtic).

17

Proposta de procés

Finalitzem aquest treball amb la proposta d'un flux de processos que permeti, amb la seva execució, obtenir el mateix resultat mitjançant la proposta exposada prèviament que el que actualment s'està aconseguint però amb una millora quant a eficiència, qualitat i amb una reducció significativa del factor risc d'error humà. La millora de l'eficiència i la reducció del factor risc d'error humà, podran ser principalment basats en el desenvolupament software plantejat, ja que, el fet d'estandarditzar els valors i donar un software de gestió que permeti canvis limitats i amb un control d'històric que doni l'opció de desfer les accions ja serà suficients per a justificar una millora clara tant en l'eficiència (per exemple, resulta més de pressa tan sols realitzar les conversions de casos que encara no hi hagin aparescut que de tots cada vegada) i la reducció del factor risc (per exemple, afegir un històric fa que, tant el treballador hagi de justificar les seves accions com que pugui veure les justificacions de canvi d'altres treballadors, fent així que pugui tenir una idea més clara del perquè de la realització de certs canvis).

Amb l'objectiu de realitzar aquesta tasca de proposta, i com ja s'ha dit en el punt d'estat de l'art, s'ha decidit basar el flux en el flux proposat per CRISP-DM. Aquesta decisió s'ha pres ja que, si és basa el procés en un DMP podrem assegurar que els resultats finals són de qualitat, ja que, els DMP resulten una manera d'assegurar que tot el treball realitzat és correcte i que per tant, aquesta correctesa es trasllada a la conclusió de l'estudi. Generant així doncs, resultats de qualitat i comprovables. A més a més, creiem que el fet de tenir un DMP amb traçabilitat (històric) permetrà que aquest sigui aplicat de manera senzilla i que a cada iteració del mateix es millori el tractament de les dades, arribant així a poder oferir un millor servei.

Així doncs, aquest apartat començarà per presentar les principals diferències entre el CRISP-DM i el procés actual per avaluar si l'impacte de canvi serà molt gran o si, simplement comportarà un canvi menor. Amb aquesta primera tasca es pretén poder detectar si el canvi serà el suficientment representatiu com per poder afectar a l'acceptació de la proposta per part dels usuaris. En tal cas, es descartarà usar CRISP-DM i es realitzarà una adaptació del flux de processos actuals al nou software. En cas contrari, es donarà una aplicació del CRISP-DM a la nostra proposta que ens permeti acreditar que, el sistema plantejat és aplicable i funcional per al cas d'estudi.

Comencem doncs per realitzar la comparativa entre l'esquema de processos final plantejat i l'esquema de processos del CRISP-DM.

17.1 Esquema processos final vs. CRISP-DM

Iniciem aquesta comparació presentant els dos gràfics de processos a comparar. En el primer gràfic mostrat (Figura 17.1) es pot veure el diagrama genèric de processos aconseguit durant l'anàlisi de procés (recordem que existeixen diagrames específics de cada subprocés que seran exposats durant la comparació) mentre que el segon és un diagrama de processos tipus CRISP-DM (Figura 17.2).

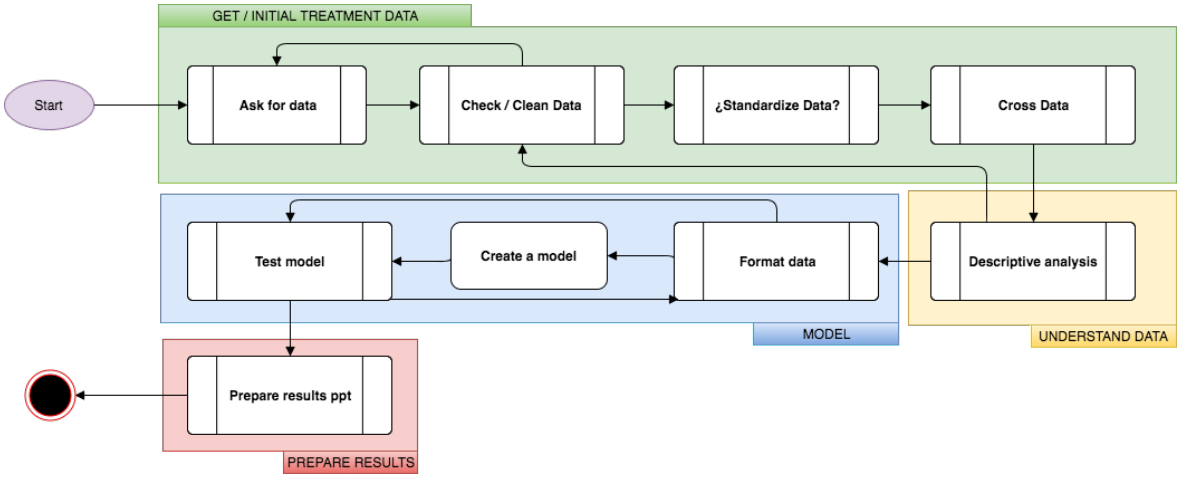


Figura 17.1: Diagrama final dels processos actuals

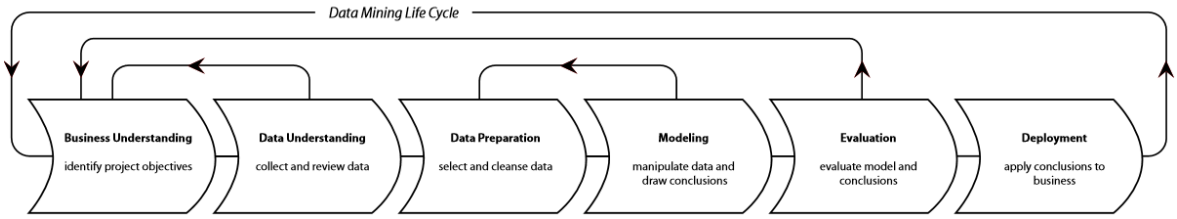


Figura 17.2: Diagrama CRISP-DM de processos

Per a realitzar aquesta comparació usarem de base el diagrama ofert per CRISP-DM, agafant les tasques d'esquerra a dreta i comparant-les amb les tasques anàlogues del diagrama de tasques actual. Així doncs es comentaran les similituds de cada una de les tasques del CRIPS-DM amb les tasques actuals i en cas de ser necessari s'entrarà en determinades subtasques per tal d'ampliar aquesta anàlisi.

Comencem doncs la comparació parlant de la tasca anomenada 'Buissness understanding' que té com a objectiu determinar l'objectiu de mercat, avaluar la situació, determinar els objectius de l'anàlisi i produir un pla de projecte. Veiem doncs que aquesta primera fase no és anàloga al diagrama de processos actual plantejat, això no és degut al fet que gran part d'aquesta tasca no es realitzi, sinó a què es va deixar fora de l'anàlisi de procés, ja que, a visió dels diferents stakeholders això no formava part del procés de tractament i estudi de les dades, sinó que resultava una capa de negoci de la qual eren més independents. Per tant, podem dir que sí que existeix una tasca actualment basada a analitzar amb el client tot el que proposa la tasca de 'Buisness Understanding' exceptuant la realització d'un pla de projecte, ja que, normalment

d'aquestes reunions tansols s'extreu una data d'entrega de l'anàlisi i posteriorment s'executa l'anàlisi amb previsió d'acabar-lo dins del termini donat. Seria un bon factor afegir aquesta planificació a la proposta de procés doncs, afegirà un control extra sobre el treball.

Seguim i ens trobem amb la tasca 'Data understanding' en la que es realitzen la recollida de dades, la descripció d'aquestes, una exploració inicial i una verificació de la qualitat. Aquesta tasca estaria composta per les subtasques 'Ask for data', 'Descriptive Analysis', 'Check / Clean Data' del diagrama de flux actual. Començant per la subtasca de 'Ask for data' veiem que, resulta encaixar amb el que es realitza en CRISP-DM, és a dir, que és una tasca en la qual bàsicament es llegeixen i s'extreuen les dades amb els mètodes més adients. Seguim amb la tasca de 'Check / Clean data', en aquest cas no encaixa amb CRISP-DM doncs com és mostra a la figura 17.3, en aquest subprocés s'està realitzant una neteja de les dades mentre que en CRISP-DM això no es proposa, per tant, probablement aquesta tasca s'hauria d'adaptar per a encabir-la en el nou diagrama de processos.

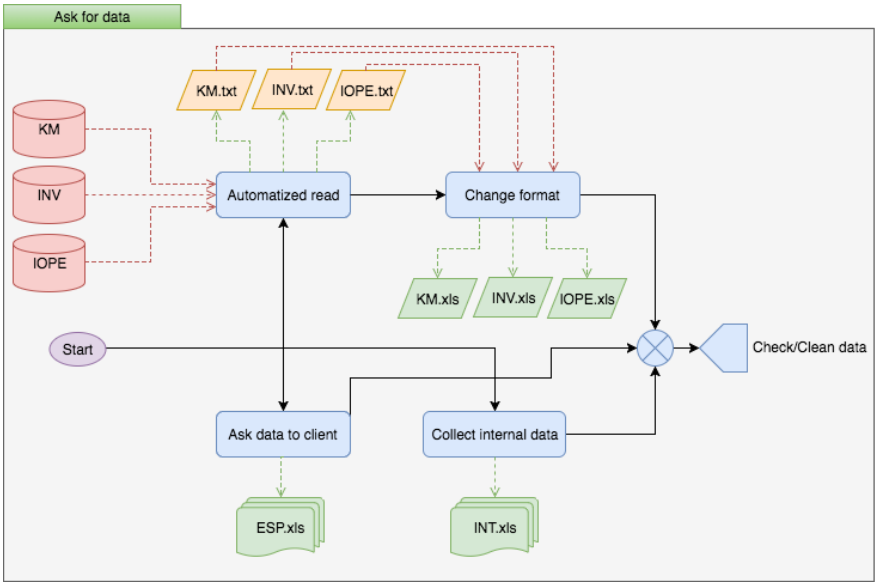


Figura 17.3: Diagrama processos: 'Demanar/Netejar dades'

Finalment i parlant ara de la tasca de 'Descriptive Analysis' veiem que, tot i que no varia en la seva definició es realitza en espais temporals molt diferents i que per tant, dóna resultats diferents, ja que, una es realitza amb les dades ja tractades (actual) i l'altre amb les dades sense tractar (CRISP-DM). Per tant, molt probablement aquesta tasca s'haurà de reiterar en cas de voler-la adaptar, doncs els resultats amb les dades netes, com podem veure en la figura X, són usats per a la presentació final de les dades. Veiem doncs que, aquesta tasca té dos objectius no complerts ni realitzats actualment, la descripció de les dades i l'avaluació de la qualitat. Una millora de la qualitat del treball podria ser la implementació d'aquests dos objectius dins de la proposta de procés final.

Passem ara a la tasca 'Data preparation' en la qual es duen a terme l'obtenció del conjunt de dades, la selecció de les dades útils, la neteja de les dades, la integració de les dades i el formateig d'aquestes. Aquesta tasca pot ser relacionada amb les subtasques 'Check/Clean Data', '¿Standardize Data?', 'Cross Data' i 'Format Data'. Trobem doncs que comencem repetint la tasca 'Check/Clean Data', ja que, com s'ha dit abans, CRISP-DM proposa realitzar la neteja de les dades en un moment diferent del que actualment s'està realitzant (figura 17.2). Pel que fa tant a la subtasca '¿Standardize Data?' com 'Cross Data' ens

trobem que fan referència a l'encreuament/integració de dades del que parla CRISP-DM i per tant, en el nostre cas específic, a l'obtenció del conjunt de dades de treball. Finalment, trobem que en aquest espai també es realitza el 'Format Data' i ens tornem a trobar en la mateixa situació que en el cas de les anàlisis descriptius doncs, resulta que aquestes tasques estan fent el mateix en ambdós casos però en diferents moments. Tot i això en aquest cas, resulta menys difícil d'adaptar-ho al funcionament actuals doncs, per tot el que s'ha exposat fins ara, sembla ser que 'Clean Data', '¿Standardize Data?', 'Cross Data' i 'Format Data' resulten ser una des agrupació de 'Data preparation' i per tant, podria ser que simplement fos l'última tasca a realitzar dins del procés de 'Data Preparation'.

Continuem parlant de la tasca 'Modelling' en la que es planteja realitzar la selecció del model, la generació dels tests, la construcció de models i l'avaluació de models. Aquesta tasca té com anàlogues 'Create Model' i 'Test Model', en aquestes tasques el que s'està realitzant resulta ser exactament el mateix que en les altres. Primerament es selecciona el model que volem usar, una vegada seleccionat es crea o es selecciona en cas de ser existent i finalment és prova i s'avalua. Per tant, aquesta tasca es realitza de manera molt similar però usant noms de i granularitats diferents.

Acabem comparant les tasques de 'Evaluation' i 'Deployment' del CRISP-DM en les que se suposa que s'avaluen els resultats, s'aproven els models, es decideixen futures accions, es preparen els resultats per mostrar i és munta el pla de seguiment. En el cas del model de procés actual es realitzen totes aquestes accions entre la 'Prepare ppt', on s'elegeixen els estudis a mostrar i per tant es seleccionen els models vàlids, es preparen per a ser mostrats, i finalment s'exposen a client (fet que no es troba dins del diagrama de processos actuals perquè com anteriorment amb la reunió inicial, s'ha suposat que és una tasca que queda fora de l'abast i sobre la que no es pot tractar directament).

Veiem doncs que en general els dos diagrames de procés presentats són molt similars i que les principals diferències són en l'ordre de realització de les tasques, en la granularitat de les tasques o en què algunes de les tasques citades per CRISP-DM no són realitzades actualment. De cara a solucionar aquestes diferències es pretén:

1. Granularitat de les tasques: ajustar les tasques per tenir la mateixa granularitat.
2. Ordre de realització de les tasques: analitzar fins a quin punt resulta rellevant l'ordre de cada tasca i plantejar o moure-la en el temps o duplicar-la.
3. Tasques no realitzades actualment: analitzar si aporten valor a l'anàlisi i plantejar integrar-les en cas afirmatiu. Val la pena estudiar l'impacte que tindríem i afegir-ho com a tasques opcionals per facilitar l'integració del projecte dins del flux de treball actual.

Així doncs creiem que no existeixen les suficients diferenciacions per a tenir un impacte rellevant sobre la implantació del projecte i que, els principals inhibidors que podem trobar poden solucionar-se fàcilment com hem enumerat anteriorment. Passem doncs a proposar el flux de treball amb una estructura basada en CRISP-DM.

17.2 Proposta de procés

La proposta de procés que realitzarem doncs es basarà en el diagrama de flux proposat per CRISP-DM amb la integració de petites modificacions per a encaixar en el cas d'estudi. Així doncs, és pretén usar totes les tasques definides per CRISP-DM, exceptuant 'Deployment' i 'Buissness Understanding' . No s'estudiarà aquests casos per complet a causa del fet que depenen totalment de l'equip directiu i les diferents negociacions que es realitzen, és a dir, s'ha de suposar que dins d'aquesta tasca existeixen unes subtasques dedicades a la negociació de preus, oferta d'anàlisi, etc. que determinaran d'alguna manera el grau de profunditat de l'anàlisi i la qualitat d'aquest. A més pressupost, més concret i més exhaustiu.

A més a més, un altre factor resulta la negociació de l'obtenció de les dades i la quantitat oferta per cada client, quedant així doncs clar que no tot dependrà de l'equip de Data Science sinó que gran part de la feina en aquestes dues tasques serà de negociació i informació per part de l'equip de ventes o directiu. Cal doncs recordar que aquest factor estarà present durant tota l'anàlisi tot i no ser-hi explícitament.

Procedim ara doncs a l'exposició del diagrama de procés que ens permetrà assegurar la viabilitat de la proposta realitzada. Per a realitzar aquest diagrama de procés s'ha decidit usar un sistema de paquets en el que, cada paquet representarà una de les tasques a realitzar. Cada paquet contindrà dins seu un conjunt de subtasques que seran explicades amb deteniment posteriorment en diferents seccions.

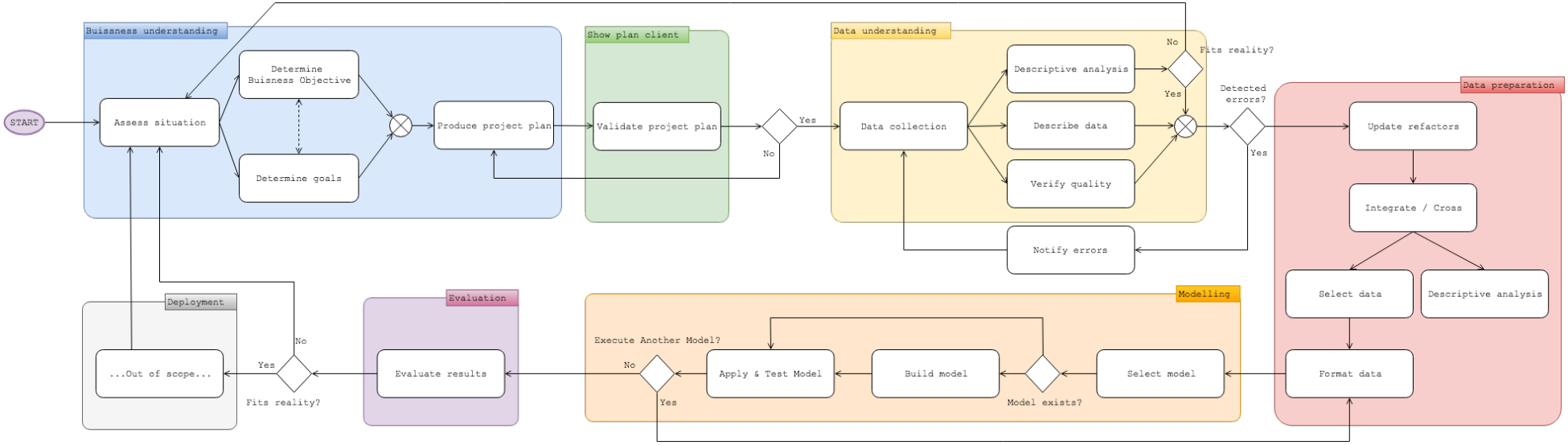


Figura 17.4: Proposició d'un nou flux de processos

17.2.1 Buisness understanding : Asess situation

Aquest resultarà ser el primer dels passos a seguir per a la realització de la feina quan es comenci de 0 una feina o a l'inici de cada iteració. Durant aquesta tasca es tractarà d'entendre el mercat en el qual treballa el client i la seva situació actual, així doncs, és de vital importància que, en aquesta fase la informació que es rep sigui sincera i real, ja que, l'escenari que és plantegi serà l'escenari des del qual l'equip de Data Science començarà. Suposem doncs que, per qualsevol motiu, es rep una informació falsa sobre la situació de la companyia. En tal cas, totes les hipòtesis i anàlisi què es realitzin no encaixaran amb la realitat i per tant, no seran vàlides.

Resulta important que en aquesta fase s'expliqui el màxim al client sobre com funciona l'estudi analític de les dades per intentar que comprengui la importància d'una comunicació directa i sincera entre l'equip i els clients. També resultarà important en aquesta tasca que, s'identifiqui els diferents representants de connexió entre les dues empreses, és a dir, els dos equips o el conjunt de persones que es comunicaran per tal de portar a terme aquest servei. Aquest fet resultarà important de cara a la tasca d'extracció de dades.

Un altre punt a tractar durant aquesta tasca serà la definició per part del client sobre les dades que és poden oferir. Aquest fet serà important de tractar, doncs depenent de la qualitat de les seves dades, és podran realitzar diferents tipus d'anàlisis. A millor quantitat i qualitat, millors anàlisis i menys cost de realització.

Aquesta tasca està plantejada per a ser realitzada durant la primera reunió de cada una de les iteracions. I acaba una vegada entes el context sobre el qual és treballarà. Tancar aquesta tasca, no representa que els diferents treballadors no segueixin formant-se sobre el sector de l'empresa sinó que, la comunicació amb el client sobre el tema s'ha tancat. Un treballador, hauria de formar-se continuadament en el sector (mirar notícies, valors, etc.) per poder adaptar els anàlisis a la realitat momentanea.

17.2.2 Buisness understanding : Determine buissness objectives

Una vegada entesa la situació de partida de la iteració, es determinaran els objectius del negoci de l'empresa, aquesta tasca serà portada a terme en paral·lel amb la determinació d'objectius de l'anàlisi donada la seva semblant naturalesa. Així doncs en aquest àmbit es tractarà de determinar quin son els objectius de millora del client, ja siguin ventes, visibilitat, valor de marca, etc. I en dependència amb aquests objectius es determinarà quins serien uns hipotètics objectius de les anàlisis per tal que aquests dos encaixin. Resulta doncs important trobar objectius de negoci clars, ja que, això afavoreix molt les anàlisis i la comprensió d'aquests. Per exemple, imaginem que un client busca incrementar les seves ventes, molt probablement resultaria un bon índex usar les ventes i la inversió diària en GRPS però si en canvi un client busca donar una visió d'exclusivitat aquest índex no resultaria òptim.

Aquesta tasca està plantejada per a ser realitzada durant la primera reunió de cada una de les iteracions. I acaba una vegada s'han determinat els objectius de negoci.

17.2.3 Buisness understanding : Determine goals

Com ja s'ha dit anteriorment aquesta tasca es realitzarà en paral·lel amb la determinació d'objectius del negoci. Així doncs, es buscaran objectius de les anàlisis relacionats amb els objectius de negoci per tal que aquests tinguin sentit. Resulta molt important determinar uns bons objectius de les anàlisis, ja que, si usem objectius no adequats els resultats no serviran per al cas d'estudi. Així doncs val la pena que es perdi temps en trobar objectius clars i útils per a cada cas d'ús.

Aquesta tasca està plantejada per a ser realitzada durant la primera reunió de cada una de les iteracions. I acaba una vegada s'han determinat els objectius de les anàlisis.

17.2.4 Buisness understanding : Produce project plan

Aquesta tasca s'activarà una vegada acabada la primera reunió de cada una de les iteracions i tractarà de realitzar un 'planning' temporal de realització de les anàlisis ofertes als clients. Amb aquesta planificació s'establiran dates límit per a cada una de les tasques a portar a terme, fet que, permetrà portar una millor gestió de qualsevol de les anàlisis. Tot i que pot no semblar rellevant de cara al client, aquesta planificació ajudarà al fet que el client s'hagi d'implicar per tal que el servei pugui ser acabat dins del termini. Sabem per tot l'estudiat dins de l'estudi de context que un dels problemes existents del tractament de les dades és que, arriben tard i en formats diferents cada vegada. Mitjançant aquesta planificació i un plec de condicions, podem arribar a fomentar que el client enviï les dades dins de termini i amb un format determinat. En cas contrari, podem justificar retrassos per adaptar el sistema a les noves dades i formats no esperats, doncs existeix una planificació temporal que ell també haurà de complir.

17.2.5 Show plan client: Validate project plan

En aquesta etapa, que no existeix i ha hagut de ser creada per aquest cas, es tractarà tan sols una sola tasca que és la validació de la planificació realitzada prèviament. En aquest cas, s'enviarà un informe al client amb la planificació temporal realitzada per veure si és correcte o necessiten canvis per adequar amb els timings. Aquesta tasca doncs pot acabar de dues maneres.

- *S'accepta el pla de projecte:* En aquest cas, s'inicia la tasca de 'Data understanding'
- *És rebutja el pla de projecte:* En aquest cas, es torna a realitzar el pla de projecte amb les modificacions demanades pel client.

17.2.6 Data understanding : Data collection

Pel que fa a aquesta tasca resulta ser de les més difícils de realitzar. Comencem doncs dient que dependrà de la planificació temporal donada en el pla de projecte, és a dir, suposem que s'ha determinat que el client haurà d'enviar a dia X les dades. Llavors el dia X, es dispararà aquesta tasca i s'haurà de realitzar tant la lectura de les dades de client mitjançant els processos SQL implementats de tipus específic com les dades globals donades per les fonts de dades estudiades mitjançant els scripts ja existents i els processos de lectura fonts de dades globals. Assegurant així que s'ha llegit totes les dades necessàries a la base de dades.

Una millora que s'ha pensat de cara a aquesta tasca és que, en comptes de realitzar-se per vàries persones de cada equip com fins ara, tan sols sigui realitzada per dues persones, una de cada equip (client i IKI Media Communications). Aquest fet facilitarà la comunicació i evitarà la duplicació de dades o els possibles errors entre fitxers font. Per això, resultava important durant el 'Buisness understanding' entendre també l'equip de treball del client per a poder determinar quin ha de ser l'interlocutor. La tria d'aquest interlocutor en un inici haurà de ser basant basada en conceptes subjectius però a mesura que s'implementi aquest mètode es pot arribar a generar un perfil de bon interlocutor basat en l'experiència per tal de realitzar aquesta selecció. Així doncs, es pretén aplicar un patró de comunicació per a l'obtenció de dades com el de la figura 17.6 en comptes d'usar l'actualment usat figura 17.5.

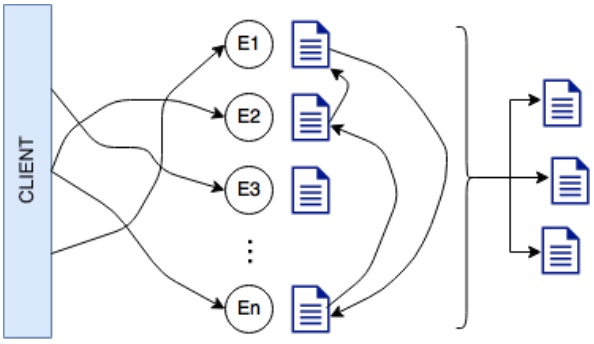


Figura 17.5: Obtenció de les dades específiques actual

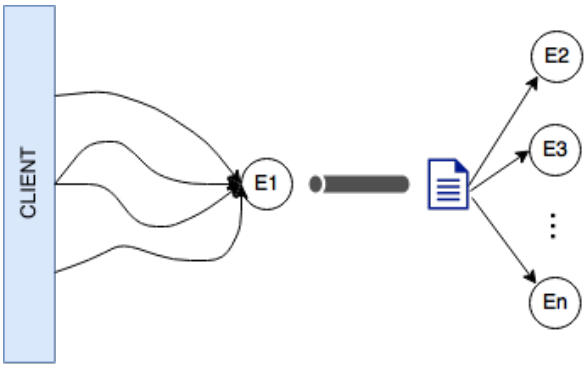


Figura 17.6: Proposta d'obtenció de les dades específiques (1 - manual)

A més a més, també es proposa un sistema per a clients molt ben preparats tecnològicament i amb confiança amb la companyia. Cal remarcar que aquesta proposta resulta bastant idíl·lica actualment, ja que, el sector publicitari és un sector antic i en general, el client no sol donar aquests tipus d'accés. Aquesta última proposta es basa a substituir les persones comunicant-se per bases de dades comunicades entre si, és a dir, que per a realitzar una obtenció de dades tan sols fes falta llençar un script que realitzes una consulta sobre la base de dades del client i que posteriorment aquesta consulta fos emmagatzemada en la base de dades local. Podem veure-ho a la figura 17.7.

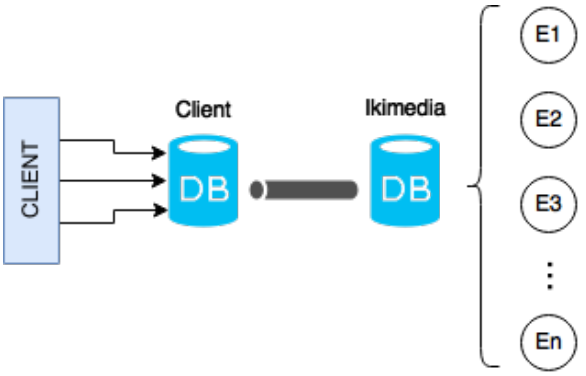


Figura 17.7: Proposta d'obtenció de les dades específiques (2 - directe)

Aquest procés s'acabarà un cop escrites totes les dades tant globals com específiques a la base de dades i actualitzades les taules de conversions amb els nous valors que hi hagin aparegut durant les lectures. Cal recordar que les lectures s'han separat en dos processos executables en moment diferents. Això ens permetrà que en cas de gran urgència sigui possible realitzar una primera lectura ràpida per treballar de manera paral·lela en la lectura i la preparació per l'encreuament. Aquest fet no es veu reflectit en el diagrama de procés, ja que, resulta ser un cas especial i s'ha decidit que era millor tan sols deixar-ho citat, doncs suposaria donar la possibilitat a trencar el flux de treball determinat i es prefereix establir un sistema de funcionament robust i estandarditzat.

17.2.7 Data understanding : Descriptive Analysis

Un cop acabada la lectura de les dades es disparen tres subtasques en paral·lel i una d'elles és l'anàlisi descriptiu inicial de les dades brutes. Com ve diu el nom de la tasca es tractarà de seleccionar les dades introduïdes a la base de dades mitjançant un script de R (sabem que R és conectable a SQL pel punt d'estudi del software) i executar una anàlisi descriptiu exhausitu. És a dir, aquest script realitzarà tots els gràfics descriptius possibles tot i que, probablement posteriorment no tots podràn ser estudiats. Aquest fet és deu a què, existeix molta memòria en els servidors i es creu que, resulta millor tenir el màxim d'informació possible i que al ser una tasca que podrà realitzar-se automàticament amb un script implementat no suposarà una càrrega de feina molt gran.

Aquest procés s'acabarà un cop extrets les anàlisis descriptius de les dades llegides. En cas de trobar desviacions molt grans amb l'escenari es tornarà a l'inici de la iteració per a intentar d'entendre el perquè les dades no encaixen amb la descripció de la situació donada. En cas contrari s'esperarà a les altres tasques i és comprovarà si hi ha hagut qualsevol error durant aquesta fase inicial de l'anàlisi. En cas afirmatiu, es notificarà de la baixa qualitat de les dades i és tornaran a demanar. Mentre que en cas negatiu es seguirà amb el flux esperat i es passarà a actualitzar les conversions existents.

17.2.8 Data understanding : Describe data

Un altre de les tasques que es dispararà serà la de descriure les dades que s'han rebut, en aquest cas, es demanarà anar actualitzant els fitxers de meta dades (explicat a l'estudi de datasources) de les fonts globals en cas de canvis i per a les dades específiques igual. En cas de ser el primer cop en rebre-les es demanarà generar un fitxer de metadades i en altre cas tan sols actualitzar-lo. Això permetrà a qualsevol persona dins l'empresa entendre-les directament sense necessitat de preguntar per cada camp i dubte existent, ja que, quedarà tot registrat al fitxer de meta dades.

Aquest procés s'acabarà un cop actualitzats o inicialitzats els fitxers de metadades. Un cop acabada la tasca s'esperarà a les altres tasques i és comprovarà si hi ha hagut qualsevol error durant aquesta fase inicial de l'anàlisi. En cas afirmatiu, es notificarà de la baixa qualitat de les dades i és tornaran a demanar. Mentre que en cas negatiu es seguirà amb el flux esperat i es passarà a actualitzar les conversions existents.

17.2.9 Data understanding : Verify quality

L'última tasca que es dispararà en paral·lel serà la de verificació de la qualitat de les dades. Per a verificar la qualitat simplement s'haurà d'actualitzar les taules del índex CBR plantejat amb la informació relativa a les noves dades rebudes i en cas de tractar-se de noves fonts de dades s'hauran d'inicialitzar. Resulta una bona idea pensar a guardar un històric dels diferents valors CBR índex al llarg del temps per tal de poder veure l'evolució de la qualitat de les diferents fonts de dades.

Aquest procés s'acabarà un cop actualitzats o inicialitzats els índexs de CBR. Un cop acabada la tasca s'esperarà a les altres tasques i és comprovarà si hi ha hagut qualsevol error durant l'anàlisi de qualitat. En cas afirmatiu, es notificarà de la baixa qualitat de les dades i és tornaran a demanar. Mentre que en cas negatiu es seguirà amb el flux esperat i es passarà a actualitzar les conversions existents.

17.2.10 Data preparation : Update Refactors

Un cop estudiades les dades i determinat que no existeixen erros passem a preparar-les per l'anàlisi. Per començar el primer que farem serà omplir les diferents taules de conversió que s'han actualitzat durant la lectura de les dades. Per fer-ho, usarem el software ofert a la proposta de solució. El que s'haurà de fer doncs és mitjançant els estàndards inicialitzats durant la lectura de les dades (dades específiques) codificar cada una de les conversions inserides (dades globals) per tal d'assegurar la possibilitat de l'encreuament. En aquest punt, perquè el primer que s'haurà de seguir el següent procediment:

1. Comprovar els valors estàndards existents i especialment els nous valors que eventualment hi hagin aparegut.
2. Comprovar les noves conversions trobades per a cada font de dades.
3. Informar-se sobre cada una de les conversions a realitzar per decidir quin valor estàndard s'ha d'assignar o si s'ha de generar un nou valor.
4. Actualitzar conversions amb explicacions dels motius de cada una per assegurar el encreuament i generar un històric de canvi que permeti justificar la qualitat de la feina feta i traçabilitat de l'error.

Una vegada assignades totes les conversions de les dades es donarà per finalitzada aquesta tasca i es passarà a realitzar l'encreuament o integració de les diferents fonts de dades a partir de les conversions obtingudes.

17.2.11 Data preparation : Integrate/Cross

En aquesta tasca es realitzarà una consulta sql que ens permeti obtenir un conjunt de dades amb totes les dades obtingudes de les diferents fonts integrades mitjançant les conversions prèviament establertes. Aquest conjunt de dades serà el conjunt de dades complet de treball, és a dir, a partir d'aquest punt, no s'haurà d'interactuar més amb el sistema de base de dades sinó que es treballarà directament amb el conjunt de dades guardat en csv o txt (depenent del gust del stakeholder). Aquesta consulta com s'ha explicat durant el treball haurà de ser realitzada directament via SQL però una millora que s'hauria d'implementar i que queda fora de l'abast d'aquest treball és oferir un sistema de consultes via front-end dedicat. Tot i això, actualment es pot oferir una consulta genèrica que ens permeti extreure el conjunt de dades complet que, tot i no ser personalitzable, donarà la possibilitat d'obtenir el conjunt de dades sense dependre de codificar una consulta sinó usant-ne una de ja predeterminada.

Tenint el conjunt de dades de treball complet, es tancarà aquest procés i es passarà a realitzar una anàlisi descriptiu i a seleccionar les dades amb les quals es vol treballar.

17.2.12 Data preparation : Anàlisi Descriptiu

Tot i poder semblar que aquesta tasca ja s'ha realitzat amb anterioritat i que no seria necessària, en el cas d'estudi resulta de gran importància poder realitzar uns bons gràfics descriptius, ja que, com s'ha

comentat durant tota l'anàlisi de context el sector publicitari encara no té del tot assumit el concepte del Data Science tot i que, actualment el seu creixement està essent molt gran. Així doncs resulta de gran utilitat tenir uns bons gràfics descriptius per a exposar, ja que, resulten fàcils d'explicar i entendre i donen un primer contacte amb el resultat que ajuden a tenir una connexió més estable amb el client. Així doncs, es realitzaran aquestes anàlisis exactament igual que en la tasca 'Data understanding : Descriptive Analysis' però amb un conjunt de dades net i exposable.

17.2.13 Data preparation : Data Selection

Un cop es té el conjunt de dades ens trobem que, sovint no resulta útil treballar amb tota la base de dades sinó que existeixen files que no són representatives pel cas o atributs que no aporten informació rellevant en l'àmbit d'estudi del client. Així doncs, en aquesta fase el que s'haurà de portar a terme és un filtre inicial del conjunt de dades, eliminant-ne les dades no rellevants per l'estudi. Un exemple pot ser, en el cas d'un estudi de les competències, eliminar totes aquelles marques del sector que no estan catalogades com a competència pel client.

Acabant amb aquesta tasca doncs obtenim una segona versió del fitxer de dades amb el que es treballarà i passem a donar format a aquestes.

17.2.14 Data preparation : Format Data

Amb les dades filtrades i eliminats tots els factors i files no rellevants, passem a donar format a les dades per tal de fer-les aptes pels diferents models a aplicar. Per tal de fer-ho tan sols s'ha d'agafar el conjunt de dades existents i canviar el format de les dades (afegint columnes, modificant els tipus, etc.) amb l'objectiu de fer-los aptes pel model. En general aquesta tasca va molt relacionada amb la de selecció del model a aplicar però s'ha decidit primer donar format, ja que, es creu que, molt possiblement una bona manera de treballar seria primerament generar forces conjunts de dades alternatius amb formats diferents de manera automatitzada (per exemple, trobar un mètode de modificació de format de dades categòriques a binàries) i posteriorment seleccionar el model i usar els conjunts generats que encaixin amb els models seleccionats.

17.2.15 Modelling : Select Model and Build Model

En aquesta tasca decidirem quins models volem aplicar a les dades a partir de l'acordat durant la fase de 'Buissness understanding'. També ens podem basar en les diferents anàlisis descriptius realitzats en cas d'haver descobert qualsevol detall extra que pugui afectar a l'estudi.

Una vegada seleccionats els models a usar, es comprovarà si existeixen implementacions prèvies o si s'ha d'implementar de 0. En cas de no existir, s'haurà d'implementar el model seleccionat. Resultarà de gran importància implementar-lo en forma de funció de R genèrica, és a dir, que un cop implementada pugui aplicar-se per a modelitzar d'altres conjunts de dades. Aquest fet ens permetrà reutilitzar la feina feta i reduir la càrrega de treball. En cas de ja ser existent, al saber que estarà implementada per a qualsevol conjunt de dades, tan sols l'haurem d'executar.

17.2.16 Modelling : Apply and Test Model

En aquest punt, una vegada es té els conjunts de dades vàlids per a cada model i els models seleccionats i implementats. S'haurà d'executar cada un d'ells i comprovar-ne els resultats. Un cop executats tots els models es decidirà si es vol seguir estenent l'estudi o si prefereix passar a avaluar els resultats finals de

l'anàlisi realitzat. En el primer cas s'haurà de tornar al pas de donar format a les dades, en canvi, en el segon es passarà al paquet de 'Evaluate'.

17.2.17 Evaluate : Evaluate Results

Finalment es passarà a realitzar una anàlisi dels resultats obtinguts pels diferents models aplicats i a preparar el conjunt de transparències d'exposició dels mateixos amb les principals conclusions de l'estudi. Aquest mètode d'exposició es proposa que sigui diferent de l'actual. Es proposa doncs que, s'ofereixin dos artefactes, un primer que serà una presentació ppt amb les conclusions i les bases principals extretes de l'estudi i un annex amb l'estudi complet realitzat justificat perquè, en cas de dubte en el procés realitzat i els resultats obtinguts, els clients puguin consultar-ho i tenir un contacte directe amb la feina feta.

Una vegada acabada aquesta feina es passarà al 'Deployment' que bàsicament serà l'execució de les decisions preses a partir de l'estudi. En aquest punt no s'han definit subtasques ni processos, ja que, a diferència dels altres, dependrà totalment de les decisions de negoci i acords realitzats.

17.3 Conclusió de la proposta de procés

Concloem doncs que s'ha donat una proposta de flux de processos que demostra que el sistema proposat és aplicable en l'àmbit i que afegeix una millora rellevant tant en la qualitat de les anàlisis com en l'estandardització del procés. Resulta rellevant també comentar que, des de l'opinió pròpia del desenvolupador, tot i que, el diagrama de procés en si no dóna una gran informació de com usar el sistema informàtic (concretament) sinó una visió més amplia i genèrica, aporta un valor molt gran, ja que, proposa un sistema de funcionament que assegurarà les bones formes.

18

Prototipus

Comencem a realitzar aquest últim punt d'implementació d'un prototipus que ens permeti demostrar alguns dels conceptes explicats durant la proposta de solució del problema plantejat. Cal recordar que aquest punt resulta ser un extra pel que fa a aquest treball que s'ha pogut donar pel fet que no han existit alteracions majors en la planificació del projecte.

Donat que el temps per implementar aquest prototip (una setmana i escaig), s'ha hagut de seleccionar un subconjunt de funcionalitats per tal de poder fer-ho assolible en el temps que es té. Així doncs, s'ha preguntat a l'equip l'opció de discutir-ho per veure quines eren les funcionalitats que més els preocupaven o a les que més valor donaven i volien veure's en funcionament. Especialment s'ha tingut en compte a l'equip directiu i a l'analista programador, pel fet que com ja s'ha vist en la figura 11.2 de l'anàlisi de stakeholders són les persones apoderades de l'empresa i per tant, a qui més rellevant resultarà convèncer.

Així doncs, i gràcies a la informació aportada. S'ha decidit que el prototip a realitzar seria tan sols basat en la font de dades de més rellevància de les globals, és a dir, Kantar Media (KM) i en una font de dades específica a seleccionar a mode d'exemple. Aquesta font de dades específica no podia ser de cap client per a poder ser mostrada fora de l'empresa (recordem que existeixen termes legals sobre les dades que es reben de client i per tant, no es poden usar les dades recollides de clients reals o resultaria ser privat de cara a persones externes de l'empresa). A causa d'això, s'ha decidit extreure aquestes dades d'una font pública de matriculacions.

Amb aquest prototipus es busca demostrar que es podran creuar les dades específiques i les globals, fet que sembla ser el principal objectiu per part dels stakeholders apoderats. Un altre punt a tenir en compte, serà mostrar de manera senzilla com es gestionarà el sistema d'històric, ja que, tot i no haver mostrat un interès directe en el tema, sembla que no s'acaba d'entendre exactament que representa i fins a quin punt resultarà útil.

S'ha obtingut un conjunt de dades de Kantar Media relacionat amb el sector de l'automòbil i informació referent a les matriculacions mensuals d'automòbils amb les que es treballarà com a font de dades específica de prova durant la implementació del prototip. La implementació d'aquest prototipus s'ha basat en quatre fases diferents.

1. Implementació de les bases de dades de les dues fonts de dades, de conversió i d'estandardització.
2. Implementació dels mètodes de lectura.
3. Implementació del back-end del software.
4. Implementació del front-end del software.

Comencem doncs a exposar part per part la feina realitzada. Amb l'objectiu de fer-ho es presentaran els scripts que permetran entendre l'estructura de base de dades preparada, es donarà una explicació breu del sistema de lectura implementat i és presentaran captures del software generat tot explicant les diferents funcionalitats que ofereix.

18.1 Implementació de les bases de dades de les dues fonts de dades, de conversio i de estandarització

Iniciem per presentar les taules de valor estàndard (esquema std) que s'han implementat, ja que, és l'única que no conté referències a altres esquemes i per tant, la primera que s'haurà d'inicialitzar. Mostrem a continuació la implementació d'aquest esquema per al prototip en el següent fragment de codi:

```
CREATE TABLE [std].[client](

    [id] INT NOT NULL IDENTITY(1,1),
    [name] VARCHAR(256)
    CONSTRAINT PK_client PRIMARY KEY ([id]),
    CONSTRAINT UNIQUE_client UNIQUE([name])

);

CREATE TABLE [std].[sector](

    [id] INT NOT NULL IDENTITY(1,1),
    [name] VARCHAR(256),
    [client_id] INT NOT NULL,
    CONSTRAINT PK_sector PRIMARY KEY ([id]),
    CONSTRAINT FK_sector_std_client FOREIGN KEY ([client_id])
       REFERENCES [std].[client]([id])

);

CREATE TABLE [std].[category](

    [id] INT NOT NULL IDENTITY(1,1),
    [name] VARCHAR(256),
    [client_id] INT NOT NULL,
    CONSTRAINT PK_category PRIMARY KEY ([id]),
    CONSTRAINT FK_category_std_client FOREIGN KEY ([client_id])
       REFERENCES [std].[client]([id])

);

CREATE TABLE [std].[product](

    [id] INT NOT NULL IDENTITY(1,1),
    [name] VARCHAR(256),
```

```
    [client_id] INT NOT NULL,
    CONSTRAINT PK_product PRIMARY KEY ([id]),
    CONSTRAINT FK_product_std_client FOREIGN KEY ([client_id])
       REFERENCES [std].[client]([id])

);

CREATE TABLE [std].[announcer](

    [id] INT NOT NULL IDENTITY(1,1),
    [name] VARCHAR(256),
    [client_id] INT NOT NULL,
    CONSTRAINT PK_announcer PRIMARY KEY ([id]),
    CONSTRAINT FK_announcer_std_client FOREIGN KEY ([client_id])
       REFERENCES [std].[client]([id])

);

CREATE TABLE [std].[brand](

    [id] INT NOT NULL IDENTITY(1,1),
    [name] VARCHAR(256),
    [client_id] INT NOT NULL,
    CONSTRAINT PK_brand PRIMARY KEY ([id]),
    CONSTRAINT FK_brand_std_client FOREIGN KEY ([client_id])
       REFERENCES [std].[client]([id])

);

CREATE TABLE [std].[model](

    [id] INT NOT NULL IDENTITY(1,1),
    [name] VARCHAR(256),
    [client_id] INT NOT NULL,
    CONSTRAINT PK_model PRIMARY KEY ([id]),
    CONSTRAINT FK_model_std_client FOREIGN KEY ([client_id])
       REFERENCES [std].[client]([id])

);
```

Seguim donant la implementació realitzada per a l'esquema de conversió. Aquest esquema tan sols té relacions amb l'esquema estàndard i per tant, aquest script seria el segon en executar. Podem trobar la implementació d'aquest esquema en el codi presentat a continuació:

```
CREATE TABLE [conv].[sector](

    [id]  INT NOT NULL IDENTITY(1,1),
    [name] VARCHAR(256) NOT NULL,
    [std_id] INT,
    [client_id] INT,

    CONSTRAINT PK_sector PRIMARY KEY ([id]),
    CONSTRAINT UNIQUE_sector_name UNIQUE([name], [client_id]),
    CONSTRAINT FK_sector_standard FOREIGN KEY ([std_id]) REFERENCES
        [std].[sector]([id]),
    CONSTRAINT FK_sector_conv_client FOREIGN KEY ([client_id])
       REFERENCES [std].[client]([id])

);

CREATE TABLE [conv].[category](

    [id]  INT NOT NULL IDENTITY(1,1),
    [name] VARCHAR(256) NOT NULL,
    [std_id] INT,
    [client_id] INT,

    CONSTRAINT PK_category PRIMARY KEY ([id]),
    CONSTRAINT UNIQUE_category_name UNIQUE([name], [client_id]),
    CONSTRAINT FK_category_standard FOREIGN KEY ([std_id])
       REFERENCES [std].[category]([id]),
    CONSTRAINT FK_category_conv_client FOREIGN KEY ([client_id])
       REFERENCES [std].[client]([id])

);

CREATE TABLE [conv].[product](

    [id]  INT NOT NULL IDENTITY(1,1),
    [name] VARCHAR(256) NOT NULL,
    [std_id] INT,
    [client_id] INT,

    CONSTRAINT PK_product PRIMARY KEY ([id]),
    CONSTRAINT UNIQUE_product_name UNIQUE([name], [client_id]),
    CONSTRAINT FK_product_standard FOREIGN KEY ([std_id])
       REFERENCES [std].[product]([id]),
    CONSTRAINT FK_product_conv_client FOREIGN KEY ([client_id])
       REFERENCES [std].[client]([id])

);
```

```
CREATE TABLE [conv].[announcer](

    [id]  INT NOT NULL IDENTITY(1,1),
    [name] VARCHAR(256) NOT NULL,
    [std_id] INT,
    [client_id] INT,

    CONSTRAINT PK_announcer PRIMARY KEY ([id]),
    CONSTRAINT UNIQUE_announcer_name UNIQUE([name], [client_id]),
    CONSTRAINT FK_announcer_standard FOREIGN KEY ([std_id])
       REFERENCES [std].[announcer]([id]),
    CONSTRAINT FK_announcer_conv_client FOREIGN KEY ([client_id])
       REFERENCES [std].[client]([id])

);

CREATE TABLE [conv].[brand](

    [id]  INT NOT NULL IDENTITY(1,1),
    [name] VARCHAR(256) NOT NULL,
    [std_id] INT,
    [client_id] INT,

    CONSTRAINT PK_brand PRIMARY KEY ([id]),
    CONSTRAINT UNIQUE_brand_name UNIQUE([name], [client_id]),
    CONSTRAINT FK_brand_standard FOREIGN KEY ([std_id]) REFERENCES
       [std].[brand]([id]),
    CONSTRAINT FK_brand_conv_client FOREIGN KEY ([client_id])
       REFERENCES [std].[client]([id])

);

CREATE TABLE [conv].[model](

    [id]  INT NOT NULL IDENTITY(1,1),
    [name] VARCHAR(256) NOT NULL,
    [std_id] INT,
    [client_id] INT,

    CONSTRAINT PK_model PRIMARY KEY ([id]),
    CONSTRAINT UNIQUE_model_name UNIQUE([name], [client_id]),
    CONSTRAINT FK_model_standard FOREIGN KEY ([std_id]) REFERENCES
       [std].[model]([id]),
    CONSTRAINT FK_model_conv_client FOREIGN KEY ([client_id])
       REFERENCES [std].[client]([id])
);
```

Un cop donades les implementacions relatives a les taules de conversió i d'estandardització passem a presentar les dues implementacions per a mantenir les dades tan globals com les dades específiques, és a dir les dades de la font Kantar Media del sector automobilístic i les dades recollides de les matriculacions d'automòbils. Presentem les taules implementades per a mantenir les dades de Kantar Media:

```
CREATE TABLE [km].[campaign](

    [id] INT IDENTITY(1,1),
    [name] VARCHAR(256),
    [client_id] INT NOT NULL,

    [sector_id] INT NOT NULL,
    [category_id] INT NOT NULL,
    [product_id] INT NOT NULL,
    [announcer_id] INT NOT NULL,
    [brand_id] INT NOT NULL,
    [model_id] INT NOT NULL,

    CONSTRAINT PK_campaign PRIMARY KEY ([id]),
    CONSTRAINT UNIQUE_campaign UNIQUE([name], [sector_id], [
       category_id], [product_id], [announcer_id], [brand_id], [
       model_id], [client_id]),

    CONSTRAINT FK_campaign_sector FOREIGN KEY ([sector_id])
       REFERENCES [conv].[sector]([id]),
    CONSTRAINT FK_campaign_category FOREIGN KEY ([category_id])
       REFERENCES [conv].[category]([id]),
    CONSTRAINT FK_campaign_product FOREIGN KEY ([product_id])
       REFERENCES [conv].[product]([id]),
    CONSTRAINT FK_campaign_announcer FOREIGN KEY ([announcer_id])
       REFERENCES [conv].[announcer]([id]),
    CONSTRAINT FK_campaign_brand FOREIGN KEY ([brand_id])
       REFERENCES [conv].[brand]([id]),
    CONSTRAINT FK_campaign_model FOREIGN KEY ([model_id])
       REFERENCES [conv].[model]([id]),
    CONSTRAINT FK_campaign_client FOREIGN KEY ([client_id])
       REFERENCES [std].[client]([id])

)

CREATE TABLE [km].[target](

    [id] INT NOT NULL IDENTITY(1,1),
    [target] VARCHAR(256) UNIQUE,

    CONSTRAINT PK_target PRIMARY KEY ([id])

);
```

```
CREATE TABLE [km].[advertisement](

    [id] INT NOT NULL IDENTITY(1,1),

    [emission_title] VARCHAR(256),
    [emission_description] VARCHAR(256),
    [emission_areas] VARCHAR(256),
    [emission_genre] VARCHAR(256),
    [date] DATE,
    [time] INT,
    [duration] INT,
    [type] VARCHAR(256),
    [format] VARCHAR(256),
    [content] VARCHAR(256),
    [context] VARCHAR(256),
    [typology] VARCHAR(256),
    [communication] VARCHAR(256),
    [promotion] VARCHAR(256),
    [pos_bloc1] INT,
    [pos_bloc2] INT,
    [pos_bloc3] INT,
    [spots_bloc1] INT,
    [spots_bloc2] INT,
    [spots_bloc3] INT,

    [campaign_id] INT,
    [channel_name] VARCHAR(256),

    CONSTRAINT PK_advertisement PRIMARY KEY ([id]),
    CONSTRAINT FK_advertisement_campaign FOREIGN KEY ([campaign_id
       ]) REFERENCES [km].[campaign]([id])

);

CREATE TABLE [km].[metrics](

    [target_id] INT,
    [advertisement_id] INT,
    [grps] FLOAT,

    CONSTRAINT PK_metrics PRIMARY KEY ([target_id], [
       advertisement_id]),
    CONSTRAINT FK_GRP_target FOREIGN KEY ([target_id]) REFERENCES [
       km].[target]([id]),
    CONSTRAINT FK_GRP_advertisement FOREIGN KEY ([advertisement_id
       ]) REFERENCES [km].[advertisement]([id])

);
```

Seguim ara ensenyat la base de dades específica d'automòbils que s'ha generat. Recordem que aquesta base de dades simula una base de dades rebuda per part del client. Podem veure la seva implementació en el codi següent:

```
CREATE TABLE [cars].[registrations] (

    [month] INT,
    [year] INT,

    brand_id INT NOT NULL,
    product_id INT NOT NULL,
    model_id INT NOT NULL,

    total_registrations INT NOT NULL DEFAULT 0,

    CONSTRAINT PK_client PRIMARY KEY ([month], [year], [brand_id],
       [product_id], [model_id]),

    CONSTRAINT FK_registrations_brand FOREIGN KEY ([brand_id])
       REFERENCES [std].[brand]([id]),
    CONSTRAINT FK_registrations_product FOREIGN KEY ([product_id])
       REFERENCES [std].[product]([id]),
    CONSTRAINT FK_registrations_model FOREIGN KEY ([model_id])
       REFERENCES [std].[model]([id])

)
```

Finalment i acabant amb la implementació de les taules de dades ens falten per afegir les taules d'usuaris i les que emmagatzemaran l'històric d'actuació. Aquesta implementació pot veure's reflectida en el codi mostrat a continuació:

```
CREATE TABLE [users].[user](

    [id] INT NOT NULL IDENTITY,
    [name] VARCHAR(256),
    [pwd] VARCHAR(256),

    CONSTRAINT PK_user PRIMARY KEY ([id]),
    CONSTRAINT UNIQUE_user_name UNIQUE([name])

);
```

```
CREATE TABLE [users].[log](

    [user_id] INT,
    [table_id] INT,
    [action] VARCHAR(256),
    [row_id] INT,
    [historic_value] VARCHAR(256),
    [timestamp] DATETIME,
    [new_value] VARCHAR(256),
    [comment] VARCHAR(256),

    CONSTRAINT FK_log_user FOREIGN KEY ([user_id]) REFERENCES [
       users].[user]([id]),
    CONSTRAINT FK_log_client FOREIGN KEY ([user_id]) REFERENCES [
       AESWEB].[conv].[client]([id])

);
```

18.2 Implementació dels mètodes de lectura

Una vegada exposades les diferents taules implementades per a emmagatzemar les dades procedim a presentar les implementacions dels mètodes que ens permetran realitzar les lectures i les actualitzacions de la base de dades. Comencem per mostrar el mètode que ens permetrà llegir les dades específiques dels automòbils. Val la pena recordar que aquestes dades també determinaran el contingut de l'esquema d'estàndards. Així doncs la implementació proposada és la següent:

```
ALTER PROCEDURE [fastReadCars] @month INT, @year INT, @file VARCHAR
   (256) AS
BEGIN;

   CREATE TABLE [CDB].[dbo].[temporal] (
         brand VARCHAR(256),
         product VARCHAR(256),
         model VARCHAR(256),
         quant INT
    );

    DECLARE @sql VARCHAR(256)
    SET @sql = 'BULK␣INSERT␣[dbo].[temporal]␣FROM␣"' + @file + '"␣
       WITH(FIRSTROW␣=␣1)'; EXEC(@sql);

    INSERT INTO [std].[brand]([name], [client_id]) SELECT t.[brand
       ], 1 FROM [dbo].[temporal] t WHERE NOT EXISTS(SELECT * FROM
       [std].[brand] s WHERE s.[name] = t.[brand]) GROUP BY t.[
       brand];
```

```
    INSERT INTO [std].[model]([name], [client_id]) SELECT t.[model
        ], 1 FROM [dbo].[temporal] t WHERE NOT EXISTS(SELECT * FROM
        [std].[model] s WHERE s.[name] = t.[model]) GROUP BY t.[
        model];
    INSERT INTO [std].[product]([name], [client_id]) SELECT t.[
        product], 1 FROM [dbo].[temporal] t WHERE NOT EXISTS(SELECT
        * FROM [std].[product] s WHERE s.[name] = t.[product]) GROUP
         BY t.[product];
    INSERT INTO [cars].[registrations]([month], [year], [brand_id],
         [model_id], [product_id])
         SELECT @month, @year, b.[id], m.[id], p.[id]
          FROM [CDB].[dbo].[temporal] t  LEFT JOIN [std].[brand] b
              ON b.[name] = t.[brand]
                                        LEFT JOIN [std].[model] m
                                           ON m.[name] = t.[model]
                                        LEFT JOIN [std].[product] p
                                            ON p.[name] = t.[
                                           product]
          GROUP BY b.[id], m.[id], p.[id];

    DROP TABLE [CDB].[dbo].[temporal];

END;
```

Veiem doncs que el funcionament del procediment és el següent:

1. Crear una taula de dades que permeti emmagatzemar temporalment les dades planes llegides.
2. Omplir les taules estàndard amb els nous valors trobats a la taula temporal per a cada atribut d'encreuament.
3. Refactoritzar la taula temporal amb els valors estàndard inserits i llegir les dades a la base de dades de cotxes.

Una vegada realitzat aquest procés s'han omplert tant les dades netes a les taules de dades específiques com els nous valors estàndard existents en l'esquema d'estàndards. Passem ara a veure com es fa pel cas de les dades de Kantar Media.

```
ALTER PROCEDURE [InitialReadKM] @client_id INT, @target_count INT,
   @file VARCHAR(256) AS
BEGIN;

    DECLARE @maximum INT;
    DECLARE @sql VARCHAR(500);
    DROP TABLE [CDB].[dbo].[temporal];
    CREATE TABLE [CDB].[dbo].[temporal] (

        [campaign] VARCHAR(256),
        [brand] VARCHAR(256),
        [model] VARCHAR(256),
        [sector] VARCHAR(256),
        [category] VARCHAR(256),
        [product] VARCHAR(256),
        [announcer] VARCHAR(256),
        [aux] VARCHAR(256),

        [channel_name] VARCHAR(256),
        [emission_areas] VARCHAR(256),

        [date] VARCHAR(256),
        [time] VARCHAR(256),
        [duration] VARCHAR(256),

        [type] VARCHAR(256),
        [format] VARCHAR(256),
        [content] VARCHAR(256),
        [context] VARCHAR(256),
        [typology] VARCHAR(256),
        [communication] VARCHAR(256),
        [promotion] VARCHAR(256),

        [pos_bloc1] VARCHAR(256),
        [spots_bloc1] VARCHAR(256),
        [pos_bloc2] VARCHAR(256),
        [spots_bloc2] VARCHAR(256),
        [pos_bloc3] VARCHAR(256),
        [spots_bloc3] VARCHAR(256),

        [emission_title] VARCHAR(256),
        [emission_description] VARCHAR(256),
        [emission_genre] VARCHAR(256)

    );

    WHILE (@target_count > 0)
    BEGIN
        SET @sql = 'ALTER TABLE temporal ADD grps' + CAST(
```

```
            @target_count AS VARCHAR(10)) + ' VARCHAR(256)'; EXEC (
            @sql);
         SET @target_count = @target_count - 1;
    END;

    SET @sql = 'BULK INSERT [temporal] FROM "' + @file + '" WITH(
       FIRSTROW = 1)'; EXEC(@sql);

    SET @maximum = (SELECT MAX(id) FROM [CDB].[km].[advertisement])
       ;
    IF @maximum IS NULL SET @maximum = 0;
    SET @sql = 'ALTER TABLE temporal ADD [id] INT IDENTITY(' + CAST
       ((@maximum + 1) AS varchar(10)) + ', 1)';
    EXEC(@sql);

    INSERT INTO [CDB].[conv].[sector]([name], [client_id]) SELECT
       DISTINCT [temporal].[sector], @client_id FROM [temporal]
       WHERE NOT EXISTS(SELECT * FROM [CDB].[conv].[sector] WHERE [
       conv].[sector].[name] = [temporal].[sector] AND [conv].[
       sector].[client_id] = @client_id);
    INSERT INTO [CDB].[conv].[category]([name], [client_id]) SELECT
        DISTINCT [temporal].[category], @client_id FROM [temporal]
       WHERE NOT EXISTS(SELECT * FROM [CDB].[conv].[category] WHERE
        [conv].[category].[name] = [temporal].[category] AND [conv
       ].[category].[client_id] = @client_id);
    INSERT INTO [CDB].[conv].[product]([name], [client_id]) SELECT
       DISTINCT [temporal].[product], @client_id FROM [temporal]
       WHERE NOT EXISTS(SELECT * FROM [CDB].[conv].[product] WHERE
       [conv].[product].[name] = [temporal].[product] AND [conv].[
       product].[client_id] = @client_id);
    INSERT INTO [CDB].[conv].[announcer]([name], [client_id])
       SELECT DISTINCT [temporal].[announcer], @client_id FROM [
       temporal] WHERE NOT EXISTS(SELECT * FROM [CDB].[conv].[
       announcer] WHERE [conv].[announcer].[name] = [temporal].[
       announcer] AND [conv].[announcer].[client_id] = @client_id);
    INSERT INTO [CDB].[conv].[brand]([name], [client_id]) SELECT
       DISTINCT [temporal].[brand], @client_id FROM [temporal]
       WHERE NOT EXISTS(SELECT * FROM [CDB].[conv].[brand] WHERE [
       conv].[brand].[name] = [temporal].[brand] AND [conv].[brand
       ].[client_id] = @client_id);
    INSERT INTO [CDB].[conv].[model]([name], [client_id]) SELECT
       DISTINCT [temporal].[model], @client_id FROM [temporal]
       WHERE NOT EXISTS(SELECT * FROM [CDB].[conv].[model] WHERE [
       conv].[model].[name] = [temporal].[model] AND [conv].[model
       ].[client_id] = @client_id);

END;
```

En aquest cas veiem que es segueix un protocol semblant a l'usat fins ara, primerament s'ha generat una taula temporal on guardar les dades i si han bolcat totes les dades. I posteriorment s'han actualitzat les taules de conversió amb els nous valors existents i trobats. La principal diferència doncs és que, en aquest cas encara no s'ha inserit cap tipus de dada de la font de dades sobre les taules especialitzades per aquesta font. Això és degut al fet que, es va decidir fer servir un sistema de doble procés que assegures que, des d'un bon inici i sense tenir totes les dades llegides, és pogués treballar en qualsevol moment sobre la taula de conversions. Aquest segon procés resulta ser el següent i bàsicament tan sols completa la font de dades km usant les conversions inserides.

```
ALTER PROCEDURE [LongReadKM] @target_list VARCHAR(256),
   @target_count INT, @client_id INT AS
BEGIN

    INSERT INTO [km].[campaign]([name], [client_id], [sector_id], [
       category_id], [product_id], [announcer_id], [brand_id], [
       model_id])
        SELECT t.[campaign], @client_id, s.[id], c.[id], p.[id], a
           .[id], b.[id], m.[id]
        FROM [dbo].[temporal] t   LEFT JOIN [conv].[sector] s ON t.
           sector=s.[name] AND s.[client_id]=@client_id
          LEFT JOIN [conv].[category] c ON t.category=c.[name] AND
             c.[client_id]=@client_id
          LEFT JOIN [conv].[product] p ON t.product=p.[name] AND p
             .[client_id]=@client_id
          LEFT JOIN [conv].[announcer] a ON t.announcer=a.[name]
             AND a.[client_id]=@client_id
          LEFT JOIN [conv].[brand] b ON t.brand=b.[name] AND b.[
             client_id]=@client_id
          LEFT JOIN [conv].[model] m ON t.model=m.[name] AND m.[
             client_id]=@client_id
        WHERE NOT EXISTS(SELECT * FROM [km].[campaign] WHERE
           sector_id = s.[id] AND category_id = c.[id] AND
           product_id = p.[id] AND  announcer_id = a.[id] AND
           brand_id = b.[id] AND model_id = m.[id])
        GROUP BY t.[campaign], s.[id], c.[id], p.[id], a.[id], b.[
           id], m.[id]

    SET IDENTITY_INSERT [CDB].[km].[advertisement] ON;

    INSERT INTO [CDB].[km].[advertisement] ([id], [campaign_id], [
       channel_name], [date], [time], [duration], [type], [format
       ], [content], [context], [typology], [communication], [
       promotion], [pos_bloc1], [spots_bloc1], [pos_bloc2], [
       spots_bloc2], [pos_bloc3], [spots_bloc3], [emission_title],
       [emission_description], [emission_genre], [emission_areas])
        SELECT  temp.[id], c.[id], [channel_name], CONVERT(date, [
           date], 103) , substring([time],1,2) * 360 + substring([
           time],4,2) * 60 + substring([time],7,2) , substring([
```

```
            duration],1,4) * 60 + substring([duration],6,2), [type],
             [format], [content], [context], [typology], [
            communication], [promotion], [pos_bloc1], [spots_bloc1],
             [pos_bloc2], [spots_bloc2], [pos_bloc3], [spots_bloc3],
             [emission_title], [emission_description], [
            emission_genre], [emission_areas]
          FROM [CDB].[dbo].[temporal] AS temp JOIN [CDB].[km].[
             campaign] AS c ON temp.[campaign] = c.[name]

    SET @target_list = @target_list + ',';
x
    DECLARE @tgt_id VARCHAR(256);
    DECLARE @sql VARCHAR(256);

    DECLARE @index INT = 1;

    WHILE (LEN(@target_list) > 0)
    BEGIN
        SET @tgt_id = SUBSTRING(@target_list, 0, CHARINDEX(',',
            @target_list));
        SET @sql = 'INSERT␣INTO␣[CDB].[km].[metrics](
            advertisement_id,␣target_id,␣grps)␣SELECT␣[id],␣' +
            @tgt_id + '␣,␣CAST(REPLACE(REPLACE(grps' + CAST(@index
            AS VARCHAR(10)) + ',␣''.'',␣''0''),␣'',''␣,␣''.'')␣AS␣
            FLOAT)␣FROM␣[dbo].[temporal]';
        EXEC(@sql);
        SET @target_list = REPLACE(@target_list, @tgt_id + ',' , ''
            );
        SET @index = @index + 1;
    END

    DROP TABLE [CDB].[dbo].[temporal];

END;
```

18.3 Implementació del front-end del software

Pel que fa al front-end i al back-end tan sols s'ensenyaran imatges presses de les diferents pantalles implementades, aquest fet és degut al fet que el funcionament del codi ha quedat explicat en la proposta de solució i per tant, ens sembla redundant. Tot i això, per a qualsevol dubte, el codi està obert en l'annex del projecte i es pot consultar. Les pantalles que s'han acabat implementant han estat les necessàries per a realitzar les conversions per part de l'usuari, per consultar l'històric i per autenticar-se. Aquestes són les mínimes necessàries per a complir amb l'objectiu proposat per a aquest prototip. Així doncs comencem exposant la pantalla de log in implementada.

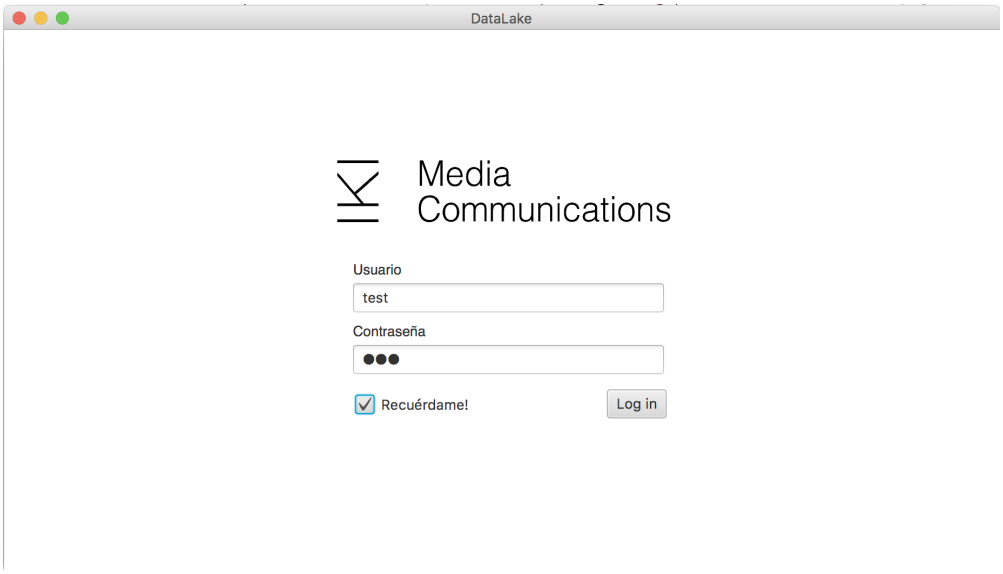


Figura 18.1: Pantalla: Log in

En la pantalla de log in mostrada a la figura 18.1, veiem que s'ha realitzat el mínim, essent això un sistema bàsic d'autenticació que avisa de si les dades introduïdes són correctes. En cas de no ser-ho, avisa de si és un error de contrasenya o un error d'usuari. Un cop autenticat l'usuari, les seves dades queden en la memòria del programa i qualsevol acció que realitzi quedarà gravada amb el seu nom. Així doncs, d'aquest log in, es passa a la següent pantalla, on l'usuari, trobarà la barra d'accés amb: la descripció reduïda de l'usuari (foto), el menú del client (logo) i el menú principal (icona).

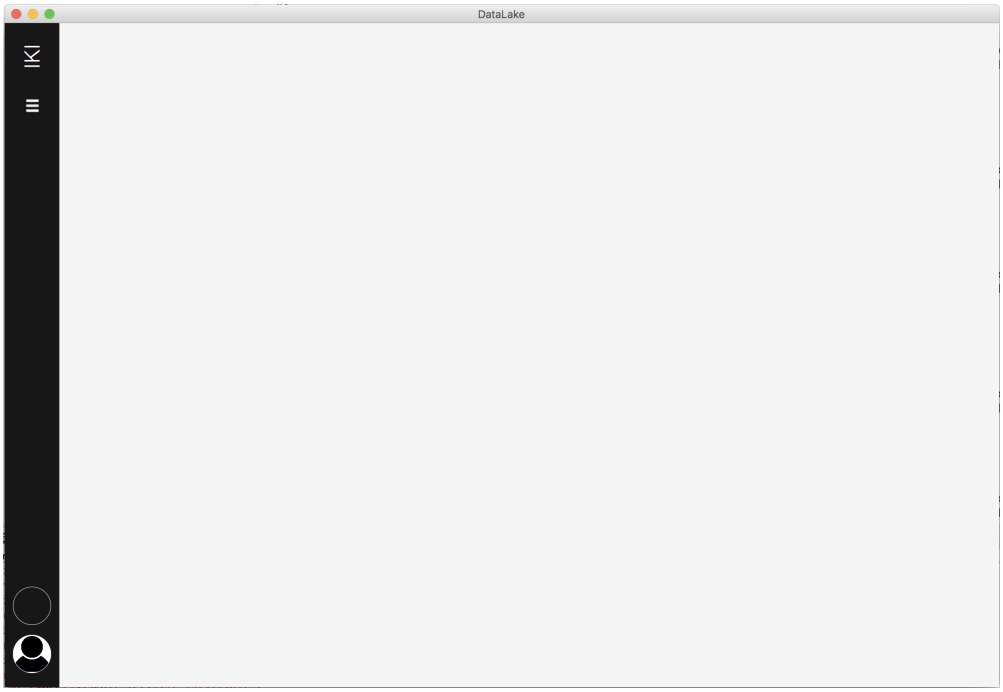


Figura 18.2: Pantalla: Base

La figura 18.2 ens mostra la primera vista que tindrem del programa, a partir d'aquesta tenim dues opcions, una primera en la qual obrirem el menú de clients per tal de seleccionar amb quin client volem treballar (sense seleccionar el client, no es podrà obrir cap de les opcions del menú principal) i una segona que obrirà el menú principal. Ambdós menús tenen diferents opcions per a seleccionar, en cas de clicar-les, es seleccionaran i el menú es tancarà. Les figures 18.3 i 18.4 presenten els dos menús oberts.

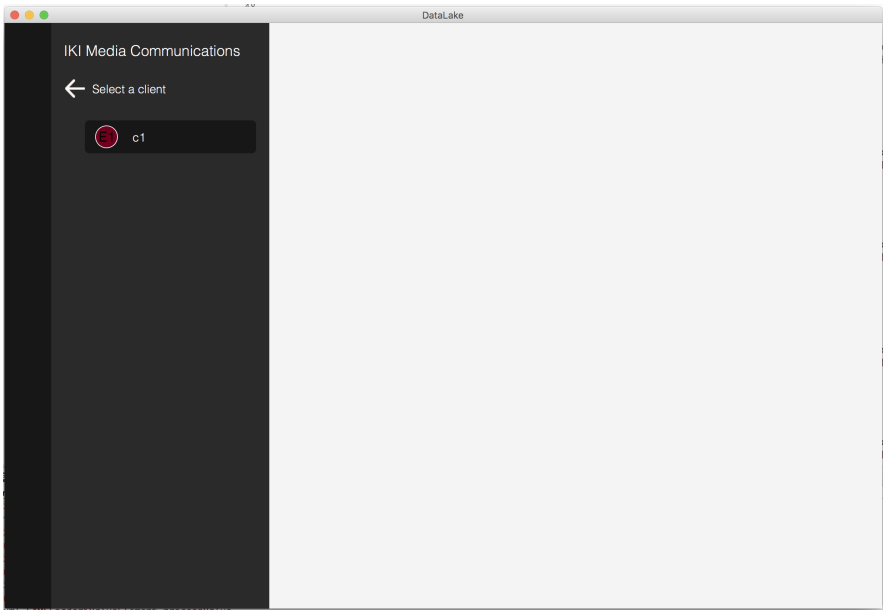


Figura 18.3: Pantalla: Menú de selecció del client

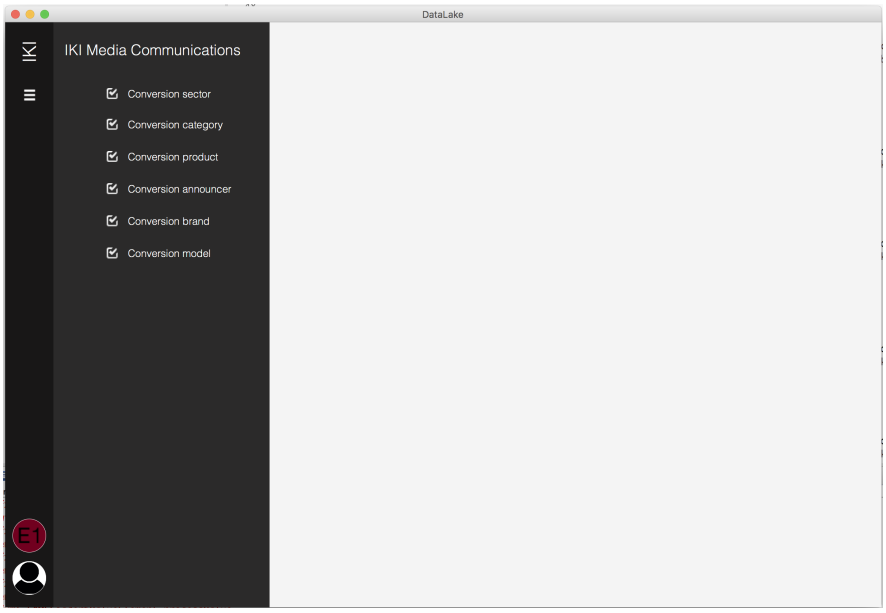


Figura 18.4: Pantalla: Menú principal

Veiem a la figura 18.2, que el logotip del client no està determinat però, en canvi, en la figura 18.4 ens apareix un logotip. Això és degut al fet que, en obrir el menú de client s'ha seleccionat una de les opcions i per tant, s'ha actualitzat la barra d'accés amb el client seleccionat. Continuem ara presentat les pantalles on es mostren les taules de conversió i les taules d'estàndard amb les seves diferents vistes. Comencem doncs presentat el cas de la pantalla d'estàndards. Recordem que, per arribar a aquesta pantalla s'ha de seleccionar una de les opcions del menú principal i un cop dins de l'opció seleccionada, triar el cas d'estàndard del menú que apareix a la nova vista.

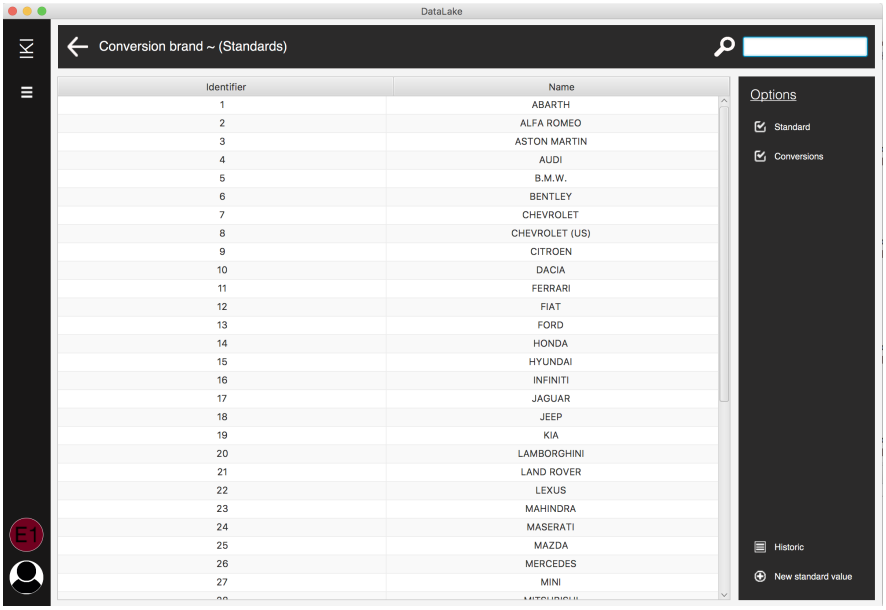


Figura 18.5: Pantalla: Taula d'estàndards

Ens trobem doncs que la pantalla representada a la figura 18.5 mostra els diferents valors estàndard que s'han determinat pel client, a més a més ofereix opcions de cerca sobre la taula i dos botons, un que ens permetrà visualitzar l'històric de creació d'estàndards i un que ens permetrà crear nous valors estàndard. Qualsevol creació d'un estàndard deixarà un registre a la taula Log, permetent així tenir un control de l'històric. Les dues finestres emergents que ens permetran controlar l'històric i crear nous estàndard són les mostrades a la figura 18.6 i 18.7.

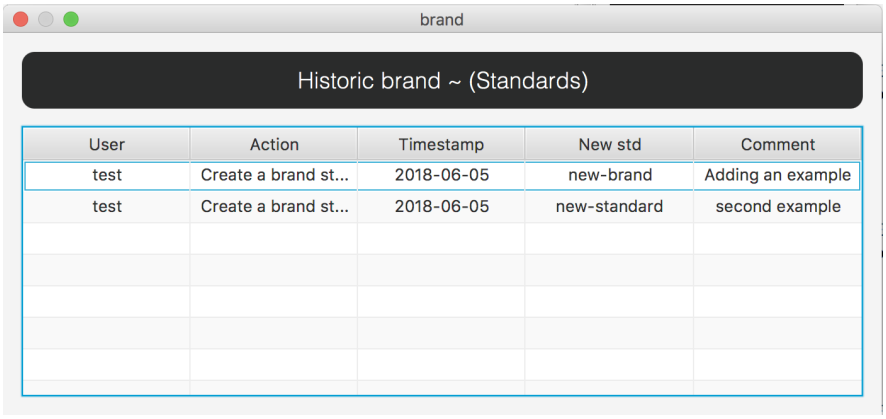


Figura 18.6: Pantalla emergent: Consultar l'històric d'estàndards

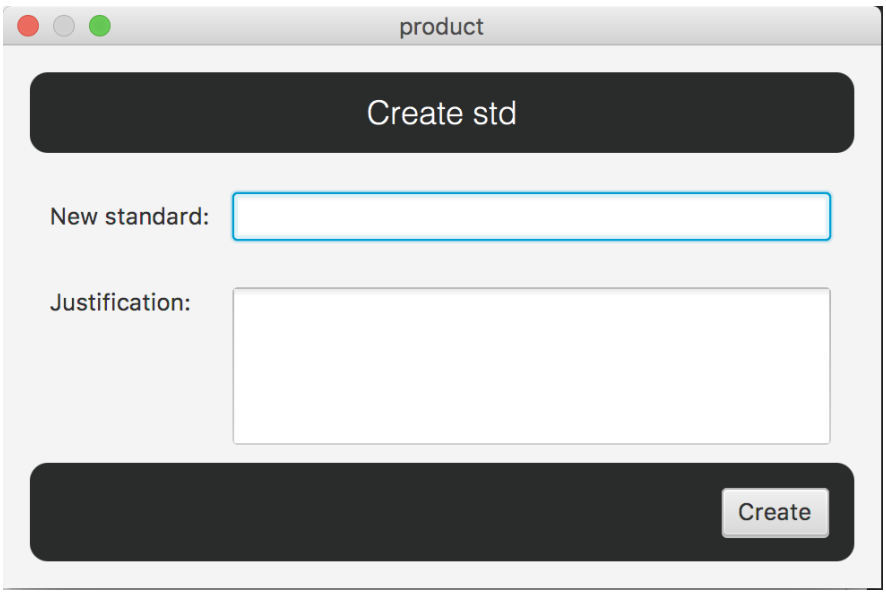


Figura 18.7: Pantalla emergent: Crear un estàndard

Passem ara al cas de la pantalla de conversions. Cal recordar que, per arribar a aquesta pantalla s'haurà d'haver seleccionat una opció al menú principal i un cop dins d'aquesta, triar el cas conversió del menú que apareix a la nova vista. És mostra la vista que ens apareixerà en la figura 18.8.

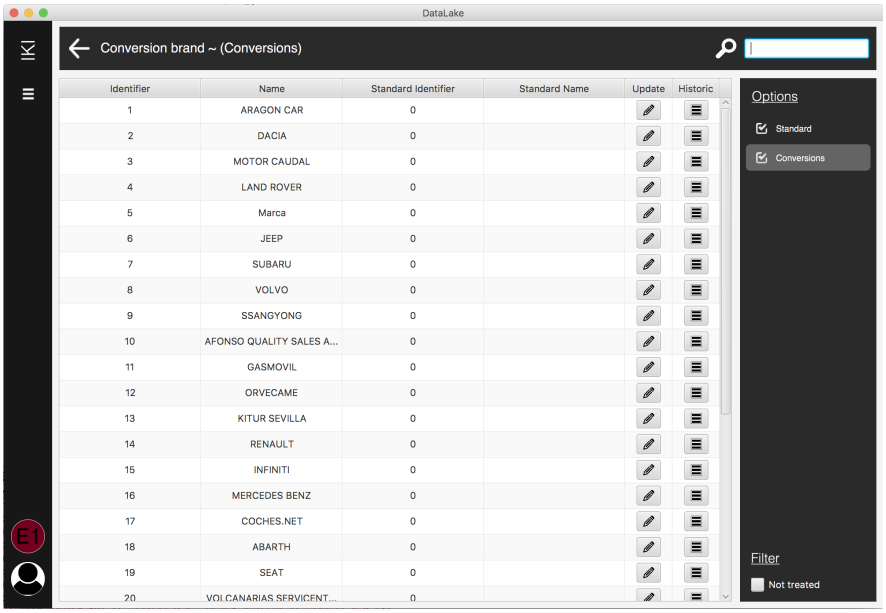


Figura 18.8: Pantalla: Taula de conversions

A la figura 18.8 podem observar que, apareix novament una taula amb les conversions existents pel client seleccionat. Aquesta taula es pot filtrar tant mitjançant la cerca per paraula com per una opció anomenada "no tractat"que eliminarà de la vista totes aquelles conversions ja establertes i, deixarà tan sols les que no tinguin un valor estàndard assignat, facilitant així el fet d'actualitzar les taules després d'una nova càrrega. A més a més, a cada fila de la taula trobem l'opció de realitzar una modificació en la conversió, és a dir, actualitzar-la i de consultar-ne l'històric. En aquest cas, l'històric resulta ser per fila, i guardarà tant el nou valor que s'ha afegit com l'antic. Podem veure les dues finestres emergents de modificació i consulta de l'històric per fila en les figures 18.9 i 18.10.

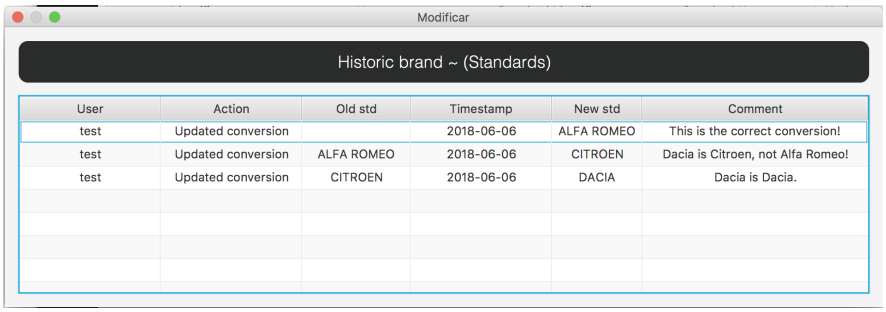


Figura 18.9: Pantalla emergent: Històric de les conversions

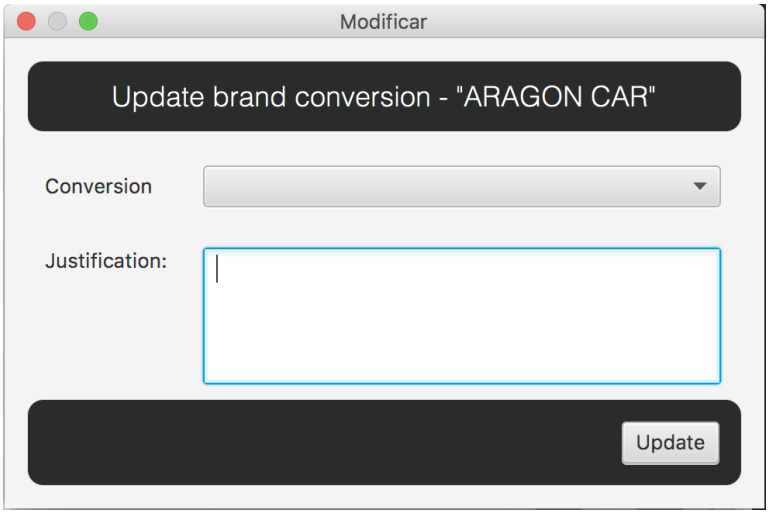


Figura 18.10: Pantalla emergent: Modificació d'una conversió

Donem doncs, amb l'exposició d'aquestes pantalles, per finalitzada la implementació de la part de front-end i com a conseqüència del nostre prototip. Així doncs, podem assegurar que mitjançant la implementació realitzada es podrà:

1. Llegir i emmagatzemar, tant la informació de Kantar Media com de la font de dades de client. (Processos de lectura MSSQL generats i taules MSSQL)
2. Tractar les conversions i els estàndards del client llegit. (Front-end i taules MSSQL)
3. Consultar l'històric d'interaccions amb el sistema. (Front-end i taules MSSQL)
4. Realitzar consultes que integrin les dades de Kantar Media i les dades de client un cop actualitzades les conversions. (Taules MSSQL). [Aquest resultat pot veure's reflectit en el fitxer de resultat d'una consulta de l'annex]

Donem doncs per vàlida la implementació del prototip en satisfer tots els requisits demanats i acordats amb les stakeholders de poder. A més a més, l'annex ofereix un vídeo en el qual es pot veure el sistema en funcionament i que permetrà entendre de manera més clara fins a quin punt de la implementació s'ha arribat.

19

Sostenibilitat : Fita final

Passem ara a omplir la taula d'anàlisi de sostenibilitat plantejada per a la realització al final del Treball de Final de Grau. S'ha decidit retirar la columna de riscos per a facilitar-ne la visualització. En cas de trobar algun risc, s'afegirà un cop exposada la taula.

	PPP	Vida Útil
Ambiental	Gairebé no existeix un impacte ambiental, ja que, tot el hardware usat és un ordinador personal i uns servidors petits d'empresa (dos, per llei). L'impacte no ha estat mesurat ja que, no es considera rellevant el resultat a tant baixa escala. S'han pres certes mesures. De fet, una de les principals restriccions inicials era no incrementar el hardware, reduint així l'impacte ambiental. No s'ha mesurat la reducció i s'opina que el màxim que podríem reduir en cas de tornar a realitzar el projecte és en material d'oficina usant directament ordinador. (A falta d'estudiar si el consum d'electricitat impacta menys que el consum de paper).	Els recursos que s'usaran durant l'ús de la proposta realitzada seran els servidors ja existents i en funcionament a l'empresa i els ordinadors dels diferents treballadors existents. El projecte no reduirà doncs l'ús actual de la tecnologia però tampoc l'incrementarà. Podem parlar doncs de què, el treball realitzat resultarà neutre quant a impacte ambiental, ja que, globalment ni l'incrementarà ni el reduirà.
Econòmica	Econòmica Els costos del projecte han estat calculats en l'apartat (6 – Costos del projecte). Per tal de reduir el cost del projecte, des d'un inici s'ha donat un marge temporal que permet solucionar errors sense incrementar ni el pressupost ni el temps. A més a més, és porta un índex de desviació en les tasques que permet reajustar-les per reduir al màxim la fluctuació de cost (és pot trobar a l'annex de seguiment dels sprints). Al final del projecte s'ha realitzat un càlcul de la diferència de hores que s'han hagut d'afegir i no resulten representatives (7-8 hores.).	El cost que tindrà el projecte durant la seva vida útil no és pot calcular. Recordem que aquest és un projecte que ha d'estar en constant desenvolupament i que aquest desenvolupament dependrà del nombre de nous clients. Tot i això, el cost, molt probablement acabarà essent el cost de la contractació d'un informàtic dedicat a la gestió i desenvolupament del sistema. És a dir, un cost baix donada la facturació de l'empresa. No s'ha tingut en compte aquest cost durant la realització del projecte ja que, no és considerat un factor rellevant al resultat baix envers la facturació.
Social	La realització d'aquest projecte ha aportat moments llargs de reflexió sobre fins a quin punt és bona idea automatitzar les coses. La discussió principal era, es pot trobar un patró de fusió de les dades que funcioni per si sol, i no hagi de ser revisat? Podrem assegurar la qualitat, sense realitzar aquesta revisió? Fins a quin punt, és útil desvincular aquest pas de la persona si potser després no tindrà tant coneixement de les dades?	El projecte soluciona el problema plantejat i ajuda a tots els stakeholders. De cara a l'analista programador, ofereix una eina per al tractament i encreuament de les dades. Aquest fet, reduirà el coll d'ampolla i els estadístics no hauran de fer tasques que no tenen assignades. D'aquesta manera també assegurem un menor risc i una major freqüència en els resultats a entregar a client. Teòricament ningú s'ha de veure perjudicat.

Taula 19.1: Taula Sostenibilitat (Fita Final)

19.1 Possibles riscos

- *Dimensió social:* un conjunt de treballadors de l'empresa poden passar a no tenir tasques a fer i com a conseqüencia quedar-se sense feina.

- *Dimensió social:* facilitar l'anàlisi de dades pot portar a fomentar l'estudi dels targets en sectors que actualment usen minoritariament el datascience facilitant dades als encarregats marketing que poden no usar èticament.

- *Dimensió ambiental:* facilitar l'anàlisi de dades pot portar a incrementar el nombre de dades a guardar i incrementar conseqüentment el consum. Tot i això, no hauria de ser rellevant durant la realització del projecte.

20

Conclusions finals

Concloem aquest projecte de final de grau assegurant que s'ha proposat un sistema software que soluciona el problema actualment existent d'obtenció i tractament de les dades. A més a més, també podem afirmar que la proposta de solució generada encaixa amb la realitat del cas d'estudi, ja que està basada en un estudi del context exhaustiu i és el màxim d'afí a aquest.

Independentment d'aquest resultat i per a justificar que aquest treball hagi estat satisfactori, es verificarà si els objectius inicials plantejats durant la definició del projecte han estat assolits. Per fer-ho, tornem a plantejar-los a continuació:

1. Estudiar el funcionament tant en l'àmbit de processos com pel que fa al software de l'obtenció i el tractament de les dades. (Estudi de context).
2. Realitzar una anàlisi complet dels requisits necessaris de la proposta donat els resultats de l'estudi previ. (Anàlisi de requisits).
3. Disseny de la proposta (Disseny del software i disseny de l'arquitectura).
4. Informe del flux de processos per a l'ús de la nova base de dades i realització de la documentació del sistema (Documentació i demostració).

Comprovem doncs com el primer objectiu ha estat realitzat, ja que, prèviament al disseny de la solució, s'ha realitzat un estudi exhaustiu del context en el qual es troba IKI Media Communications S.L.. Aquest estudi s'ha basat en la realització: d'una anàlisi de hardware, d'una anàlisi de software, d'un estudi de les fonts de dades, d'una anàlisi de stakeholders i d'un estudi de procés actual. Així doncs s'han documentat i tingut en compte tots els factors que s'han cregut rellevants de cara al disseny de la solució. Amb tot això podem concloure que el primer objectiu ha estat assolit.

Pel que fa al segon objectiu podem veure que també ha estat complert perquè, una vegada realitzat l'estudi de context s'han especificat un conjunt de requisits que hauria de complir la solució per tal d'assegurar que complirà amb tot el necessari per a oferir una millora suficient en el context estudiat. Aquests requisits per tant, ens han donat uns objectius mínims a assolir amb la proposta.

En relació al tercer objectiu afirmem que també ha estat assolit, ja que, un cop estudiat el context i fixat un conjunt de requisits a complir s'ha portat a terme la proposta d'una solució. Iniciant per la selecció del software a usar i del hardware necessari i seguint per la realització de la proposta en si de la solució. A més a més, s'ha comprovat durant aquesta solució que encaixes tant amb els requisits plantejats com amb el context estudiat, justificant cada una de les decisions presses d'acord amb conceptes prèviament estudiats o requisits establerts. Es pot comprovar doncs que durant el transcurs de la proposta, la importància que ha tingut la informació recopilada durant el treball previ ha estat molt elevada. Creiem que aquest fet és rellevant de remarcar, ja que ens permet dir que, el projecte realitzat serà útil en el context en el qual s'haurà d'aplicar i que, a més, al complir amb els requisits demanats, oferirà la solució al problema demanada.

Acabant ara amb el quart objectiu comprovem que s'ha realitzat adequadament, ja que, existeix un plantejament de procés basat en el model de DMP CRIPS-DM i perquè la documentació relativa al projecte ha quedat reflectida durant tot el projecte en aquest treball i tots els artefactes usats han estat tant comentats en el mateix com oferts en l'annex. A més a més, es vol donar especial rellevància a aquests dos àmbits. De cara al flux de processos donat, es creu que, de cara a la millora del servei, poden arribar a resultar tant o més importants que el software ofert. Pel que fa a la documentació, se li vol donar una especial importància, ja que, es creu que de cara a un equip de treball l'existència d'una documentació completa facilitarà molt el treball en un futur. Per exemple, en el cas de voler generar un API d'obtenció de les dades emmagatzemades, la documentació de taules UML resultarà de gran ajuda.

Confirmem amb aquestes justificacions que s'han complert tots els objectius plantejats per a aquest treball de final de grau. Aquest fet ja seria suficient però a més a més, com que no hi ha hagut de impediments durant la realització del projecte, tal com està explicat a la planificació, s'ha pogut implementar un prototip inicial que ens permetés veure de manera més clara la proposta realitzada. Aquest prototipus ha estat realitzat amb un subconjunt dels requisits plantejats i ha estat provat amb un cas de prova extern als clients de l'empresa.

Pel que fa al prototip, tal com hem vist, ha estat implementat i ens ha donat l'opció de veure en funcionament el software i realitzar la prova de concepte. Aquesta proba pot veure's reflectida tant en el vídeo del funcionament de l'annex com en els resultats obtinguts de l'encreuament que també podem trobar a l'annex. Donem doncs per vàlid i suficient el prototip implementat i en valorem molt el resultat final donat el temps ajustat amb el qual es comptava per a la seva realització.

Tanquem aquest projecte afirmant doncs que s'han complert tots els objectius establerts i que s'ha pogut realitzar la implementació d'un prototip com a prova de concepte i per tant, que el projecte ha resultat satisfactori i ofereix un valor real de cara a l'empresa IKI Media Communication S.L..

20.1 Valoració personal

Afegeixo aquest punt per a poder donar el punt de vista de l'estudiant sobre treball realitzat, ja que, crec que resulta important destacar el valor dels coneixements usats i apresos durant aquest treball de final de grau. M'agradaria doncs començar citant les principals assignatures que han estat útils per a la realització d'aquest treball obtinguts durant la carrera.

- *Enginyeria de Requisits (ER):* D'aquesta assignatura s'han usat tots els coneixements apresos d'anàlisi de requisits, d'estudi de context i de planificació de projectes d'enginyeria del software.

- *Gestió de Projectes del Software (GPS):* S'han usat les nocions relacionades amb la metodologia àgil i la gestió de projectes (Diagrames de Gantt, Pertt, etc.).
- *Arquitectura del Software (AS + IES):* D'aquestes dues assignatures d'especialitat s'han usat els conceptes de disseny UML adquirits i els diferents patrons de disseny.
- *Disseny de bases de Dades (DBD + BD):* Pel que fa a aquesta matèria, s'han aplicat tots els continguts adquirits de bases de dades relacionals i d'índexs d'aquestes mateixes.
- *Mineria de dades (MD):* En relació amb aquesta assignatura s'han utilitzat els conceptes obtinguts de Data mining processes i de tractament de datasets.
- *Sistemes d'informació per a Organitzacions (SIO):* Aquesta assignatura ens ha permès realitzar anàlisis de processos exhaustius i l'avaluació d'aquests.

A més a més, s'ha usat l'experiència obtinguda en els projectes de Projecte d'Enginyeria del Software (PES) i de Projecte Aplicat d'Enginyeria (PAE) per a la realització d'aquest treball de final de grau.

Vull seguir aquesta opinió personal donant una petita reflexió de la importància de la feina feta dins del món de l'enginyeria del software. Aquest fet és deu a què sovint, tant els enginyers informàtics com tots els altres professionals, valorem més obtenir un resultat què planificar i dissenyar les solucions. Això, és molt notable en el nostre sector al ser intangible (el programari no es pot tocar) però en canvi, en altres casos és impensable, per exemple, ningú construeix una casa sense el treball previ d'un arquitecte en els plànols i crec que aquest concepte s'ha d'aconseguir instaurar en el món de la informàtica per què, en cas que no tan sols es demanessin resultats sinó també bones formes i validacions, s'acabaria obtenint programari de major qualitat i seríem molt més productius.

Concloc doncs deixant la meva opinió com a enginyer del software i la meva visió d'aquest projecte. En aquest projecte he intentat donar valor a aquesta planificació i a l'estudi inicial realitzant tot el procés de disseny que ens permetrà implementar la solució en un futur de forma segura i assegurant-ne la qualitat.

Bibliografia

[1] E. Acuña, *Preprocessing in Data Mining*, pp. 1083–1085. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

[2] J. Hönigl and J. Küng, "Obtaining a data quality index with respect to case bases," *Vietnam Journal of Computer Science*, vol. 2, pp. 47–56, Feb 2015.

[3] J. Deighton, "Rethinking the Profession Formerly Known as Advertising," *Journal of Advertising Research*, vol. 57, no. 4, 2017.

[4] I. M. C. S.L, "IKI Media Communications," 2015.

[5] R. Thompson, "Stakeholder Analysis," 2018.

[6] A. L. Mendelow, "Environmental Scanning–The Impact of the Stakeholder Concept," *International Conference on Information Systems*, 1981.

[7] Z. Abdallah, L. Du, and G. Webb, *Data Preparation*. United States: Humana Press, 2016.

[8] J. Johanesson, "Why data quality programmes are important for advancement," *Journal of Education Advancement and Marketing*, vol. 3, 2016.

[9] G. Williams and Z. Huang, "Modelling the kdd process a four stage process and four element model," 07 1996.

[10] Ó. Marbán, G. Mariscal, and J. Segovia, "A Data Mining & Knowledge Discovery Process Model," *Data Mining and Knowledge . . .* , no. February, pp. 1–17, 2009.

[11] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The KDD process for extracting useful knowledge from volumes of data," *Communications of the ACM*, vol. 39, no. 11, pp. 27–34, 1996.

[12] P. Chapman, R. Kerber, J. Clinton, T. Khabaza, T. Reinartz, and R. Wirth, "The CRISP-DM process model," *The CRIsP–DM Consortium*, vol. 310, no. C, p. 91, 1999.

[13] U. Fayyad, G. Piatetsky-shapiro, P. Smyth, and T. Widener, "The kdd process for extracting useful knowledge from volumes of data," *Communications of the ACM*, vol. 39, pp. 27–34, 1996.

[14] A. Azevedo and M. Filipe Santos, "Kdd, semma and crisp-dm: A parallel overview," *Vietnam Journal of Computer Science*, pp. 182–185, 01 2008.

[15] M. Gualtieri and R. Curran, "The Forrester Wave™: Big Data Predictive Analytics Solutions, Q2 2015," *Forrester Research*, pp. 1–18, 2015.

[16] Alteryx, "Data Preparation Tools," 2018.

[17] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel, "Knime: The konstanz information miner," in *Data Analysis, Machine Learning and Applications* (C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker, eds.), (Berlin, Heidelberg), pp. 319–326, Springer Berlin Heidelberg, 2008.

[18] M. Mannion and B. Keepence, "SMART requirements," *ACM SIGSOFT Software Engineering Notes*, vol. 20, no. 2, 1995.

[19] J. Rasmusson, *The Agile Samurai–How Agile Masters Deliver Great Software*. 2010.

[20] KantarMedia, "Instar analytics," 2017.

[21] KantarMedia, "About kantar media," 2017.

[22] KantarMedia, "About kantar media," 2017.

[23] InfoAdex, "Infoadex," 2017.